



WTSCNet: A CSI feedback network based on wavelet transform convolution and attention mechanism

Yong-Li An^{a,b,*}, Shuo-Yang Lu^{a,b}, Yi-Xiao Liu^{a,b}, Ying-Chao Liu^{a,b}, Qiang Kong^{a,b}^a College of Artificial Intelligence, North China University of Science and Technology, Tangshan, 063210, China^b Hebei Key Laboratory of Industrial Intelligent Perception, Tangshan, 063210, China

ARTICLE INFO

Publishing editor: Yu-Lian He

Keywords:

CSI feedback
Frequency division duplex
Massive MIMO systems
WTConv

ABSTRACT

In order to enhance the performance of massive multiple-input multiple-output (MIMO) systems, the downlink channel state information (CSI) must be efficiently compressed and fed back to the base station in a frequency-division-duplex mode. The deep-learning-based CSI feedback networks can learn more complicated patterns in the data as compared with the traditional compressed-sensing techniques. Nevertheless, many of the current methods are founded on the convolutional neural network (CNN) modules that were initially developed to act on images and might not make full use of relevant channel and spatial information. The suggested model, wavelet transform convolution and spatial channel self-attention-channel network (WTSCNet), is a combination of wavelet-transform convolution and attention mechanisms. The encoder also has access to wavelet-transform convolution to generate multi-resolution features, enabling it to perform better in both space information and local details at a reduced computational cost. On the decoder part, it is presented as a spatial channel self-attention-channel reconstruction block (CARBlock) which combines multi-scale residual reconstruction with spatial-channel attention to improve the recovery of CSI. Experiments prove that WTSCNet enhances the indoor reconstruction precision by 6.41 dB, compared with CRNet, at a compression ratio of 1/4. In comparison with the Transformer-based TransNet+, WTSCNet requires about 9.9 million floating-point operations less than TransNet+ but offers a 0.28-dB improvement in the indoor normalized mean squared error (NMSE) at the same compression ratio. The simulation results indicate that WTSCNet offers a good accuracy-complexity trade-off between CSI compression and feedback.

* Corresponding author.

E-mail addresses: anyl@ncst.edu.cn (Y.-L. An), lsyleaon@stu.ncst.edu.cn (S.-Y. Lu), yxliu@stu.ncst.edu.cn (Y.-X. Liu), liuyc@ncst.edu.cn (Y.-C. Liu), kq3515@stu.ncst.edu.cn (Q. Kong).<https://doi.org/10.1016/j.jnlest.2026.100367>

Received 4 March 2026; Received in revised form 1 July 2026; Accepted 12 July 2026; Available online 9 September 2026.

1674-862X/© 2026 The Authors. Publishing Services provided by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1 Introduction

The multiple-input multiple-output (MIMO) technology is an essential part of next-generation communication systems, which can provide significant gains in capacity and spectral efficiency (SE) without the use of extra bandwidth. Nevertheless, the implementation requires a large number of antennas, which add a significant channel state information (CSI) feedback overhead. To achieve the maximum antenna gain and reduce interference, the base station (BS) must perform efficient precoding, a procedure that is highly sensitive to precise CSI. In frequency-division-duplex (FDD) systems, channel reciprocity does not exist, so the BS uses downlink CSI, estimated and fed back by the user equipment (UE). Thus, effective compression of the CSI matrix at the UE and its precise reconstruction at the BS is a key issue.

Traditional finite-feedback schemes are mainly based on codebook and compressed-sensing (CS) procedures. Codebook-based methods use feedback information as a binary index of a specific codeword in a predetermined codebook. Nevertheless, the codebook size is increasing exponentially with the number of antennas in the massive MIMO system, which makes it more difficult to search and store codebooks. The least absolute shrinkage and selection operator (LASSO) [1], block-matching and three-dimensional filtering approximate message passing (BM3D-AMP) [2], and total-variation minimization by augmented Lagrangian and alternating direction algorithms (TVAL3) [3] are examples of CS-based methods that rely on CSI sparsity and typically require iterative reconstruction. These can be computationally costly and have high latencies, making them unsuitable for large-scale, low-latency CSI feedback.

Following the successful implementation of deep learning (DL) in various fields, many DL-based approaches have been designed to carry out wireless communication tasks [4,5]. A number of early studies used DL to implement CSI feedback in massive MIMO systems [6]. In 2018, CsiNet was proposed by Wen et al. [7], which presented a convolutional neural network (CNN)-based autoencoder to compress and reconstruct CSI and showed that DL could be used to acquire the intrinsic structure of CSI. The next step in this direction was taken by Guo et al. [8] who suggested CsiNet+ which enhanced the performance of CsiNet by modifying the size of the convolutional kernel. However, merely augmenting the kernel size increases the computational load, and a constant kernel cannot be effectively used to model the relationships between CSI at various scales. Therefore, Lu et al. [9] presented CRNet with a multi-resolution structure consisting of different convolutional kernel sizes to enhance the reconstruction of CSI.

Attention mechanisms have also been introduced to boost the feature representation. Attention-CsiNet [10] is an extension of the CsiNet framework that utilizes both recurrent and attention modules to enhance the performance of CSI recovery. CLNet was proposed by Ji et al. [11], which simultaneously processes the real and imaginary parts of the CSI and employs lightweight attention to maintain the structural properties. The criss-cross attention network (CCA-Net) [12] also added criss-cross attention to enhance the interaction between features and maintain a low complexity. Although these approaches enhance spatial or channel feature models, striking a balance between the reconstruction accuracy and UE-side computation cost is still challenging.

Transformer-based architectures have a positive effect on long-range dependency learning. CsiFormer [13], TransNet [14], and TransNet+ [15] show that global attention is effective in the feedback of CSI but their computational costs may not be desirable in resource-limited UEs. More recent lightweight and hybrid structures, such as lightweight CNN-based feedback systems, MetaFormer-based networks, and CsiMobileViT, continue to investigate the trade-off between reconstruction quality and complexity of deployment [16–18]. The application of spatial-frequency learning based on wavelets has also proven valuable in more general applications, such as flare removal, remote-sensing restoration, and intelligent-transportation image enhancement [19–21], which implies that multi-resolution wavelet representation can be applied more generally than CSI feedback.

In this paper, we introduce the wavelet transform convolution and spatial channel self-attention-channel network (WTSCNet) as a means of compressing and providing feedback on CSI. In contrast to traditional CNN-based models with manually designed multi-resolution branches, WTSCNet uses wavelet-transform convolution (WTConv) to learn low- and high-frequency CSI characteristics at the cost of encoder complexity. The spatial and channel synergistic attention (SCSA) module is coupled with CRBlock to produce the spatial channel self-attention-channel reconstruction block (CARBlock) on the decoder side, which enhances multi-scale spatial-channel feature reconstruction.

The main contributions of this paper are summarized as follows:

- Our encoder is based on WTConv, which substitutes human-designed multi-resolution convolutional branches with wavelet-domain decomposition. The encoder has both low-frequency structural components and high-frequency localized details so that it can capture multi-scale angular-delay-domain CSI features with low UE-side computational complexity.
- Our approach is to develop a CARBlock-based decoder which combines spatial-channel synergistic attention and CRBlock-based multi-scale residual reconstruction. The design models all three types of spatial dependencies, inter-channel relationships, and local multi-scale structures, enhancing the feature-refinement ability of the BS-side decoder.
- Our contribution is the design of an asymmetric encoder-decoder structure where multi-resolution information preservation is carried out prior to UE-side compression whereas spatial-channel feature refinement is applied after BS-side codeword extension. When the compression ratio is 1/4, WTSCNet enhances the indoor normalized mean squared error (NMSE) by more than 0.28 dB compared with TransNet+ and minimizes the computational cost by about 27.8%, indicating a good reconstruction accuracy-complexity trade-off.

In contrast to CRNet, WTSCNet substitutes manual design of multi-resolution convolutional branches with wavelet-domain decomposition. It is even stronger due to CARBlock which helps enhance the fusion of spatial-channel features on the decoder side. Compared with TransNet+, WTSCNet aims to achieve a better accuracy-complexity trade-off at lower compression rates using a lighter UE-side encoder.

2 Proposed method

2.1 CSI feedback system model

In this study, we consider a single-cell FDD massive MIMO orthogonal frequency-division multiplexing (OFDM) system. The BS is equipped with N_t transmit antennas, while the UE is equipped with N_r receive antennas, where $N_t \gg N_r$. For simplicity, we set $N_r = 1$, and the system operates over N_c subcarriers. The received signal at the UE on the n -th subcarrier is expressed as

$$y_n = \tilde{\mathbf{h}}_n^H \tilde{\mathbf{v}}_n x_n + w_n \quad (1)$$

where $\tilde{\mathbf{h}}_n \in \mathbb{C}^{N_t \times 1}$, $\tilde{\mathbf{v}}_n \in \mathbb{C}^{N_t \times 1}$, $x_n \in \mathbb{C}$, and w_n denote the downlink channel vector, precoding vector, transmitted symbol, and additive noise, respectively. The superscript $(\cdot)^H$ denotes the Hermitian (conjugate-transpose) operation. The frequency-domain downlink CSI matrix is expressed as

$$\tilde{\mathbf{H}} = \begin{bmatrix} \tilde{h}_{0,0} & \tilde{h}_{0,1} & \cdots & \tilde{h}_{0,N_t-1} \\ \tilde{h}_{1,0} & \tilde{h}_{1,1} & \cdots & \tilde{h}_{1,N_t-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{h}_{N_c-1,0} & \tilde{h}_{N_c-1,1} & \cdots & \tilde{h}_{N_c-1,N_t-1} \end{bmatrix} \in \mathbb{C}^{N_c \times N_t} \quad (2)$$

Considering massive MIMO systems with large N_c and N_t , direct feedback requires $2N_c N_t$ real values for the real and imaginary channel components. To reduce this overhead, the channel matrix is transformed into the angular-delay domain, as follows:

$$\mathbf{H} = \mathbf{F}_d \tilde{\mathbf{H}} \mathbf{F}_a^H \quad (3)$$

where $\mathbf{H}$ denotes the full angular-delay-domain CSI matrix, and $\mathbf{F}_d \in \mathbb{C}^{N_c \times N_c}$ and $\mathbf{F}_a \in \mathbb{C}^{N_t \times N_t}$ are discrete Fourier transform (DFT) matrices along the delay and angular dimensions, respectively. Owing to the limited delay spread, most channel energy is concentrated in the first N_c delay bins. Retaining these rows yields the truncated angular-delay-domain CSI $\mathbf{H}_a \in \mathbb{C}^{N_c' \times N_t}$.

The encoder compresses $\mathbf{H}_a$ into a low-dimensional codeword:

$$\mathbf{v} = f_E(\mathbf{H}_a; \boldsymbol{\theta}_E) \quad (4)$$

where $f_E(\cdot; \boldsymbol{\theta}_E)$ denotes the encoder mapping parameterized by the trainable parameter set $\boldsymbol{\theta}_E$. The semicolon separates the network input from its parameter set. The codeword is transmitted to the BS through the feedback link. The decoder reconstructs CSI as

$$\hat{\mathbf{H}}_a = f_D(\mathbf{v}; \boldsymbol{\theta}_D) \quad (5)$$

where $f_D(\cdot; \boldsymbol{\theta}_D)$ denotes the decoder mapping parameterized by the trainable parameter set $\boldsymbol{\theta}_D$. Therefore, the complete feedback process is

$$\hat{\mathbf{H}}_a = f_D(f_E(\mathbf{H}_a; \boldsymbol{\theta}_E); \boldsymbol{\theta}_D) \quad (6)$$

The objective is to jointly learn $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_D$ such that the reconstruction error between $\mathbf{H}_a$ and $\hat{\mathbf{H}}_a$ is minimized.

2.2 Overall architecture of WTSCNet

This section discusses the design of WTSCNet and its components. Fig. 1 shows the general encoder-decoder structure. The encoder reduces the CSI to a low-dimensional codeword and the decoder recovers the CSI at the BS. The encoder has been designed to support multi-scale CSI structures but with low computational complexity because information that is lost at the UE cannot be retrieved by the decoder.

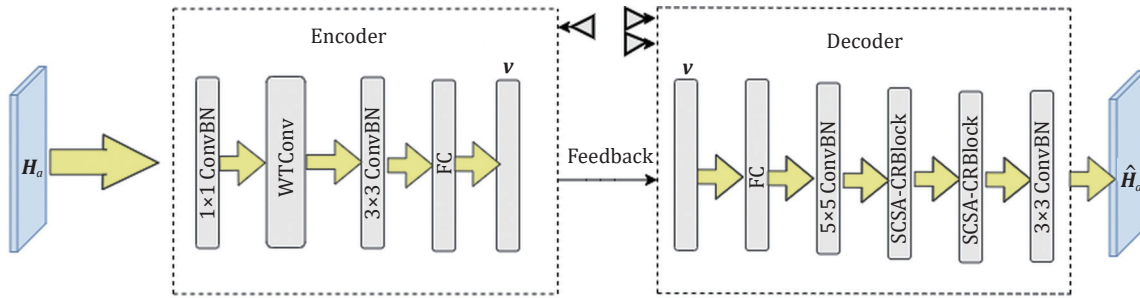


Fig. 1. Overall architecture of WTSCNet. Here ConvBN denotes convolution followed by batch normalization; FC denotes a fully connected layer.

CSI in the angular-delay domain consists of localized sparse multipaths and structures at various scales. The short-time Fourier transform (STFT) has a fixed window and so gives a fixed time-frequency resolution. The discrete cosine transform (DCT) and fast Fourier transform (FFT) employ global basis functions and are useful as a global spectral representation or energy compaction; however, both of them cannot localize the features by scale and position. Conversely, the wavelet transform (WT) gives a coefficient $c_{j,k} = \langle \mathbf{X}, \psi_{j,k} \rangle$, with $\mathbf{X}$ being the input signal or feature map, $\psi_{j,k}$ being the wavelet basis function at scale j and translation index k , and $\langle \cdot, \cdot \rangle$ denotes the inner product. Therefore, the coefficient is scaled and indexed jointly by scale and position, allowing localized multi-resolution representation. Table 1 shows these differences.

Table 1

Comparison of transform-domain representations.

Transform	Spatial localization	Multi-resolution	Main characteristic
FFT	No	No	Global frequency representation
DCT	No	No	Energy compaction with global bases
WT	Yes	Yes	Localized structure and detail subbands

In order to take advantage of these properties, WTSCNet adds WTConv to the encoder to obtain CSI features at various resolutions. CARBlock on the decoder uses attention with residual reconstruction to encode the spatial, channel, local, and global information. The purpose of this design is to enhance the trade-off between the reconstruction accuracy and computational complexity.

2.3 WTConv-based encoder

The input CSI first passes through a complex-valued input layer that organizes the real and imaginary parts into two feature channels. A 1×1 convolution adjusts the feature-channel dimension without changing the spatial resolution. WTConv [22] is then used in place of a conventional multi-resolution branch, followed by a 3×3 convolution and a fully connected layer that produces the compressed codeword $\mathbf{v}$.

The 1×1 convolution reduces or expands the feature channels and controls the computational load of subsequent layers. WTConv uses wavelet-domain decomposition to capture the low- and high-frequency information at different scales. The WTConv structure is shown in Fig. 2.

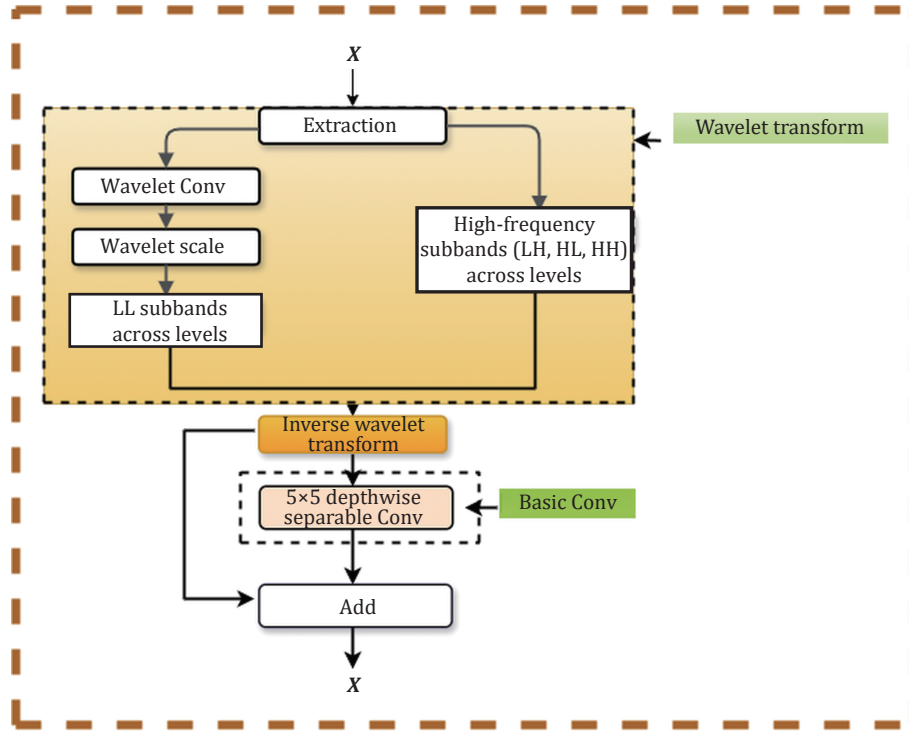


Fig. 2. Structure of WTConv. Here the LL-subband branch stores $\mathbf{X}_{LL}$ components across wavelet decomposition levels, whereas the high-frequency branch stores the corresponding $\mathbf{X}_{LH}$, $\mathbf{X}_{HL}$, and $\mathbf{X}_{HH}$ components.

Let $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ denote the input feature map, where C , H , and W denote the number of channels, height, and width, respectively. The two-dimensional discrete wavelet transform (DWT) decomposes $\mathbf{X}$ into the low-low (LL), low-high (LH), high-low (HL), and high-high (HH) subbands:

$$[\mathbf{X}_{LL}, \mathbf{X}_{LH}, \mathbf{X}_{HL}, \mathbf{X}_{HH}] = \mathcal{W}(\mathbf{X}) \quad (7)$$

where $\mathcal{W}(\cdot)$ denotes the DWT operator. Subband-wise convolution is expressed as

$$\tilde{\mathbf{X}}_s = \phi_s(\mathbf{X}_s), \quad s \in \{LL, LH, HL, HH\} \quad (8)$$

where s is the subband index and $\phi_s(\cdot)$ is the convolutional transformation for subband s . The processed feature map $\mathbf{Y}$ is reconstructed using the inverse WT, as follows:

$$\mathbf{Y} = \mathcal{W}^{-1}(\tilde{\mathbf{X}}_{LL}, \tilde{\mathbf{X}}_{LH}, \tilde{\mathbf{X}}_{HL}, \tilde{\mathbf{X}}_{HH}) \quad (9)$$

The low-frequency subband contains the main CSI structure, whereas the high-frequency subbands maintain the direction and local changes. The next 3×3 convolution improves the local characteristics and the fully connected layer performs the final dimensionality reduction.

WTConv is unlike regular downsampling using pooling, because the high-frequency subbands are not thrown away until feature fusion occurs. The low- and high-frequency elements are processed and then reconstructed back into the original form by applying inverse transformation. It also minimizes the likelihood of the loss of localized sparse multipath information in WTConv due to irreversible damage. However, the ultimate mapping to a low-dimensional codeword necessarily deletes some information, which is even worse when the compression ratio is high.

2.4 CARBlock-based decoder

The decoder first expands the compressed codeword through a fully connected layer. A 5×5 convolution extracts larger-scale local features, after which two sequential CARBlocks refine the feature representation. A final 3×3 convolution reconstructs the desired CSI matrix. CARBlock combines the SCSA module [23] with CRBlock, as shown in Fig. 3.

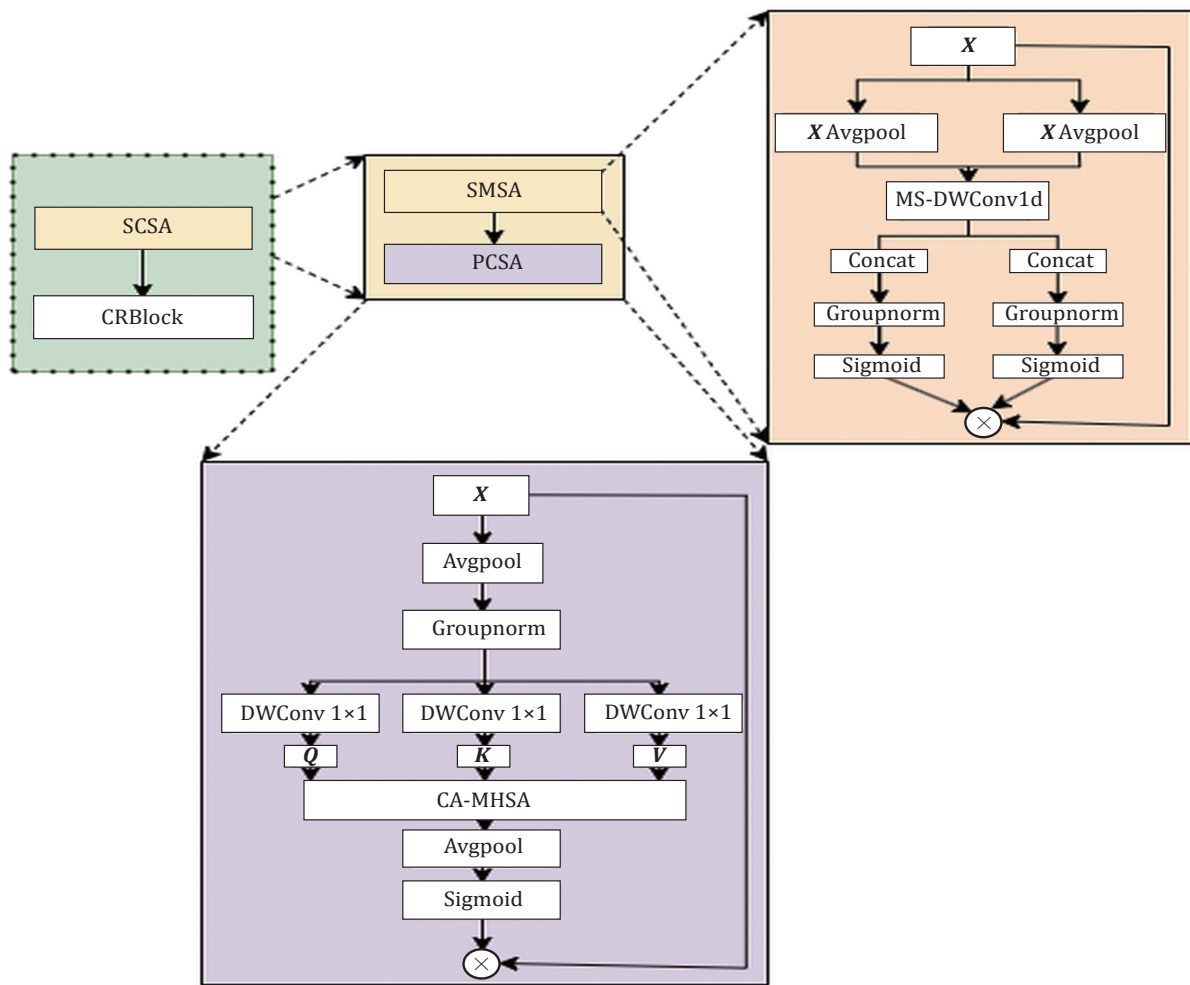


Fig. 3. Structure of CARBlock. Here SMSA denotes shareable multi-semantic spatial attention; PCSA denotes progressive channel-wise self-attention; MS-DWConv1d denotes multi-scale depth-wise one-dimensional convolution; DWConv denotes depth-wise convolution; CA-MHSA denotes channel-wise multi-head self-attention.

The combination of spatial and channel attention in SCSA allows it to improve the process of feature extraction and model long-range interactions. The spatial branch uses average pooling in the height and width directions to extract position-related information. Following this, local and global convolutions can be used to exploit spatial relationships at various levels and a sigmoid-activated attention map can highlight

informative CSI areas. The channel branch generates queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) via downsampling and convolution and assigns higher weights to informative channels with the help of self-attention. CRBlock also extracts multi-scale features using various convolutional paths and retains the input using residual connections. Spatial, channel, and multi-scale information fusion of CSI reconstruction is enabled by combining SCSA and CRBlock.

3 Experiments

This section defines the simulation conditions and assesses WTSCNet in terms of reconstruction accuracy, complexity, generalization, robustness, and communication performance.

3.1 Simulation environment settings

The COST2100 channel model [24] was used to generate indoor and outdoor massive MIMO channel datasets. A uniform linear array with $N_t = 32$ antennas and half-wavelength spacing is deployed at the BS. After the two-dimensional DFT, only the first 32 delay-domain rows were retained, resulting in a 32×32 angular-delay-domain CSI matrix. Let M denote the length of the compressed codeword. The compression ratio is defined as $\eta = M / (2N_c N_t)$, where $2N_c N_t$ is the number of real-valued elements in the truncated complex CSI matrix. The model was trained and tested using PyTorch on an NVIDIA RTX 3090 graphics card (GPU). Both the training and test sets had 100000 and 30000 samples, respectively. The training was performed for 1000 epochs using a batch size of 200. Adam was used with an initial learning rate of 2×10^{-3} , and cosine annealing reduced the learning rate to 5×10^{-5} .

The mean squared error (MSE) is used as the training loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{H}}_{a,i} - \mathbf{H}_{a,i}\|_{\text{F}}^2 \quad (10)$$

where N denotes the mini-batch size, i is the sample index, and $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm. Reconstruction performance is primarily evaluated by the NMSE in decibels:

$$\text{NMSE}_{\text{dB}} = 10 \lg \mathbb{E} \left[\frac{\|\hat{\mathbf{H}}_a - \mathbf{H}_a\|_{\text{F}}^2}{\|\mathbf{H}_a\|_{\text{F}}^2} \right] \quad (11)$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator. A more negative NMSE indicates a better reconstruction. The CSI correlation coefficient ρ and cosine similarity were also reported where required. The computational complexity is reported in terms of floating-point operations (FLOPs).

3.2 Simulation results analysis

WTSCNet was compared with representative CNN-, attention-, and Transformer-based CSI feedback networks under indoor and outdoor scenarios, as shown in Table 2.

For both indoor and outdoor scenarios, WTSCNet achieves a lower NMSE than CsiNet, CRNet, CLNet, and CsiNet+. At $\eta = 1/4$, WTSCNet achieves -33.40 dB indoors and -16.00 dB outdoors, improving CRNet by 6.41 dB and 3.30 dB, respectively. Compared with TransNet+, WTSCNet improves NMSE by 0.28 dB indoors and 0.20 dB outdoors while reducing FLOPs from 35.650M to 25.754M, corresponding to a reduction of approximately 27.8%. Fig. 4 illustrates the NMSE-complexity comparison.

WTSCNet does not consistently outperform TransNet+ at high compression ratios. When the codeword becomes extremely short, a large amount of global channel information must be represented by a limited number of latent variables. Transformer-based models can explicitly model long-range dependencies and therefore retain an advantage in several 1/16 to 1/64 cases. WTSCNet largely enhances local multi-resolution extraction by WTConv and spatial-channel fusion with CARBlock. Thus, the performance at extreme compression may be enhanced by a simple global modeling module or entropy-constrained latent representation.

Table 2

Comparison of NMSE and FLOPs between WTSCNet and existing networks.

Method	η														
	1/4			1/8			1/16			1/32			1/64		
	FLOPs	NMSE (dB)		FLOPs	NMSE (dB)		FLOPs	NMSE (dB)		FLOPs	NMSE (dB)		FLOPs	NMSE (dB)	
		Indoor	Outdoor		Indoor	Outdoor		Indoor	Outdoor		Indoor	Outdoor		Indoor	Outdoor
CsiNet	5.140M	-17.36	-8.75	4.730M	-12.70	-7.61	3.840M	-8.65	-4.51	3.580M	-6.24	-2.81	3.450M	-5.84	-1.93
CRNet	5.120M	-26.99	-12.70	4.070M	-16.01	-8.04	3.550M	-11.35	-5.44	3.280M	-8.93	-3.51	3.160M	-6.49	-2.22
CLNet	4.050M	-29.16	-12.88	3.010M	-15.60	-8.29	2.480M	-11.15	-5.56	2.220M	-8.95	-3.49	2.090M	-6.34	-2.19
CsiNet+	24.570M	-27.37	-12.40	23.520M	-18.29	-8.72	23.000M	-14.14	-5.73	22.740M	-10.43	-3.40	22.610M	-6.32	-3.27
Attn-CSI	24.720M	-20.29	-10.43	22.620M	/	/	21.580M	-10.16	-6.11	21.050M	-8.58	-4.57	20.790M	/	/
TransNet	35.720M	-32.38	-14.86	34.700M	-22.91	-9.99	34.140M	-15.00	-7.82	33.880M	-10.49	-4.13	33.750M	-6.08	-2.62
TransNet+	35.650M	-33.12	-15.80	34.600M	-23.47	-9.86	34.080M	-15.70	-7.88	33.820M	-11.98	-4.87	33.690M	-7.99	-3.07
WTSCNet	25.754M	-33.40	-16.00	24.705M	-20.39	-9.786	24.180M	-14.34	-6.277	23.919M	-10.34	-3.807	23.788M	-7.024	-2.37

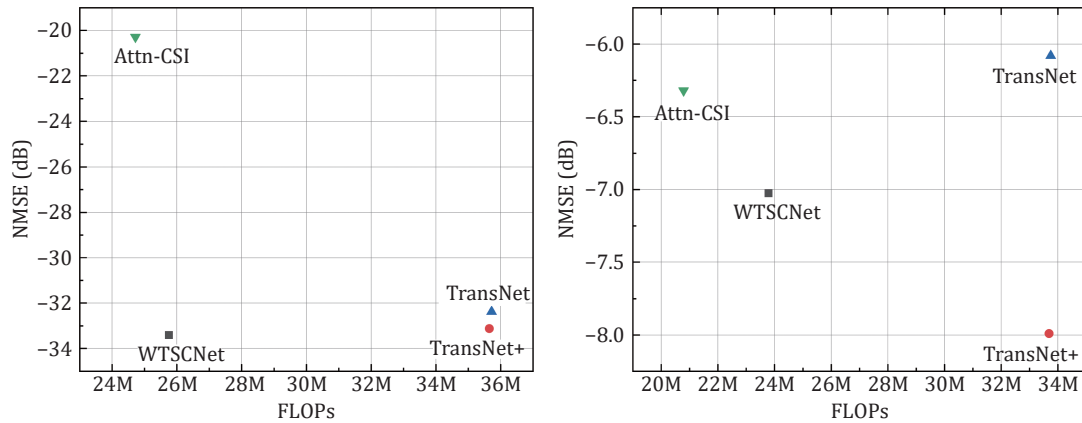
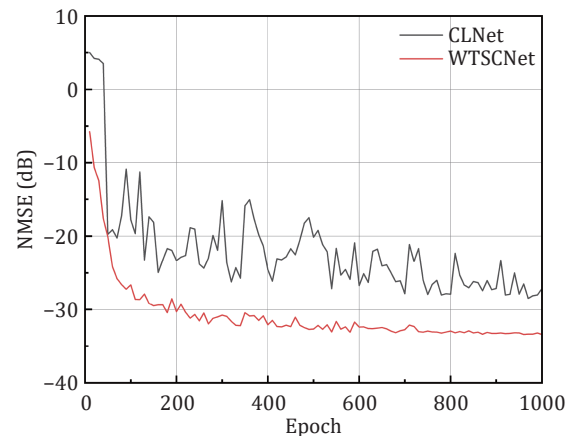
**Fig. 4.** NMSE-FLOPs comparison at compression ratios of (a) 1/4 and (b) 1/64.

Fig. 5 compares CLNet and WTSCNet under the cosine-annealing learning-rate schedule. WTSCNet achieves a lower NMSE over the same training period and exhibits a relatively smooth convergence trend.

3.3 Analysis of generalizability and training stability

The original mixed-scenario experiment combined indoor and outdoor COST2100 samples and compared WTSCNet with CLNet. The results are shown in Table 3.

WTSCNet consistently outperforms CLNet under mixed-scenario training. The corresponding NMSE gains are 3.450 dB, 2.550 dB, 1.900 dB, 0.881 dB, and 0.800 dB from $\eta = 1/4$ to $\eta = 1/64$, respectively. At lower compression ratios, the gain is more significant since the latent representation still has enough capacity to allow the proposed multi-resolution and spatial-channel processing to be carried out successfully. The persistent benefit in all five compression ratios also suggests that the suggested feature-extraction method is functional even with mixed indoor and outdoor channel samples.

**Fig. 5.** Indoor NMSE curves at a compression ratio of 1/4.

To distinguish mixed-scenario training from direct transfer to an unseen propagation environment, Table 4 reports the source-to-target evaluations at $\eta = 1/4$. SE at 10 dB was included as a communication-level indicator.

The strict source-to-target experiment is a zero-shot out-of-distribution test in which there are no examples of the target condition that can be trained on. Both transfer directions show significant performance decline, which suggests that there is a large domain difference between the indoor and outdoor COST2100 channels, which can be quantified in terms of delay spread, angular sparsity, multipath geometry, and power distribution. Hence, the existing training approach has not yet resulted in a completely domain invariant CSI representation.

The matched-scenario, mixed-scenario, and strict cross-scenario protocols have tested various deployment assumptions. Matched-scenario outcomes consider reconstruction when the training and deployment settings are similar, but mixed-scenario outcomes determine whether one model can train on heterogeneous channel samples that are simultaneously available during training. Conversely, the strict cross-scenario experiment tests direct deployment on an entirely unknown propagation setting. They indicate that any zero-shot deployment would involve target-domain calibration, domain adaptation, or domain-robust representation learning.

The training stability was evaluated using three independent random seeds in the indoor scenario at $\eta = 1/4$. To directly characterize the run-to-run variability, the NMSE degradation relative to the best repeated run was reported, while the correlation coefficient and cosine similarity were presented in their absolute forms. As shown in Table 5, the maximum relative NMSE degradation among the three repeated runs is 1.5737 dB, with a standard deviation of 0.8341 dB. Although the exact reconstruction error varies moderately across different random initializations, the correlation coefficient and cosine similarity remain highly consistent, with mean values of 0.996681 and 0.999585 and standard deviations of only 1.14×10^{-4} and 7.60×10^{-5} , respectively. These results indicate that WTSCNet maintains stable structural and directional consistency in the CSI reconstruction across repeated training runs.

3.4 Ablation experiment

To verify the effectiveness of CARBlock and WTConv, we replace CARBlock with CRBlock and WTConv with ordinary convolution. The corresponding variants are denoted as WTSCNet-CR and WTSCNet-Conv, respectively. Table 6 presents the results at $\eta = 1/4$.

Replacing CARBlock with CRBlock degrades NMSE by 2.46 dB and 0.88 dB for indoor and outdoor scenarios, respectively, quantitatively confirming the contribution of spatial-channel attention. Reverting WTConv into the conventional form of convolution increases NMSE by 2.06 dB and 1.22 dB in indoor and outdoor environments, respectively, which supports the application of wavelet-domain multi-resolution feature extraction. The two ablations affect various steps of the feedback pipeline: WTConv enhances the representation before codeword compression, and CARBlock enhances the feature refinement after codeword expansion. Their complementary benefits mean that the

Table 3
NMSE (dB) comparison between WTSCNet and CLNet using mixed datasets.

Method	η				
	1/4	1/8	1/16	1/32	1/64
WTSCNet	-19.60	-13.60	-9.30	-6.481	-4.40
CLNet	-16.15	-11.05	-7.40	-5.600	-3.60

Table 4
Strict cross-scenario evaluation of WTSCNet at $\eta = 1/4$.

Training scenario	Test scenario	NMSE (dB)	ρ	Cosine	SE at 10 dB (bit/s/Hz)
Indoor	Outdoor	0.2049	0.2752	0.1078	0.8381
Outdoor	Indoor	0.8889	0.3701	0.1771	1.2335

Table 5
Multi-seed stability of WTSCNet where Δ NMSE is the measured value relative to the best repeated run.

Seed	Δ NMSE (dB)	ρ	Cosine
3407	0.0000	0.996809	0.999671
2025	1.2664	0.996641	0.999558
7777	1.5737	0.996592	0.999526

Table 6
Ablation experiment at $\eta = 1/4$.

Method	NMSE (dB)	
	Indoor	Outdoor
WTSCNet-Conv	-31.34	-14.78
WTSCNet-CR	-30.94	-15.12
WTSCNet	-33.40	-16.00

ultimate performance cannot be brought about by just one component, but by the coordinated encoder-decoder structure.

The increased indoor gain observed with CARBlock indicates that spatial-channel refinement is especially useful when the angular-delay structure is not so spread out. In an outdoor setting, WTConv and CARBlock have more similar effects, as expected on the basis of the less concentrated propagation structure. Although these observations depend on the scenarios, the complete WTSCNet is the best in both situations, and it can be confirmed that none of the modules are excessive.

The effect of the wavelet basis was further evaluated using the Daubechies-1 (db1), Daubechies-2 (db2), and Coiflet-1 (coif1) wavelets. To isolate the basis choice from the original main results, Table 7 reports the NMSE difference relative to db1.

The db1 wavelet provides the best NMSE among the tested bases, outperforming db2 and coif1 by 0.263 dB and 0.286 dB, respectively. Nevertheless, all three bases maintain $\rho > 0.9967$ and cosine similarity above 0.9996. Therefore, changing the wavelet family slightly affects the reconstruction-error magnitude but has little influence on the structural consistency of the recovered CSI. The result also shows that the advantage of WTSCNet is not restricted to a single wavelet basis, although db1 is the most suitable choice among the evaluated alternatives.

Table 7

Effect of different wavelet bases in the indoor scenario at $\eta = 1/4$.

Basis	ΔNMSE relative to db1 (dB)	ρ	Cosine
db1	0.000	0.996809	0.999671
db2	0.263	0.996729	0.999653
coif1	0.286	0.996727	0.999650

3.5 Robustness to practical feedback impairments

The practical digital CSI feedback consists of finite-precision codewords and a non-ideal feedback channel. Thus, we consider three exemplary impairment classes: Finite-bit latent quantization, independent feedback-bit errors, and additive white Gaussian noise (AWGN) on the latent codeword. The experiments tested the ability of WTSCNet to preserve the CSI structure before converting the compressed representation back into an operational digital feedback signal.

For a latent value z , the uniform b -bit quantization over a fixed interval $[z_{\min}, z_{\max}]$ is expressed as

$$q_b(z) = \text{round}\left(\frac{\text{clip}(z, z_{\min}, z_{\max}) - z_{\min}}{z_{\max} - z_{\min}} (2^b - 1)\right) \quad (12)$$

where $q_b(\cdot)$ denotes the b -bit quantizer, $z_{\min}$ and $z_{\max}$ are the lower and upper clipping bounds, $\text{round}(\cdot)$ rounds its argument to the nearest integer, and $\text{clip}(\cdot)$ limits its argument to the specified interval. For the feedback-link AWGN, the received latent codeword is modeled as

$$\tilde{\mathbf{v}} = \mathbf{v} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}) \quad (13)$$

where $\mathbf{n}$ is the additive noise vector, $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian distribution, σ_n is the noise standard deviation, and $\mathbf{I}$ is the identity matrix. Table 8 summarizes the results for the indoor scenario at $\eta = 1/4$. For the digital feedback experiment, p denotes the independent probability that each feedback bit is flipped.

In the case of latent quantization, 8-bit and 6-bit representations can still maintain correlation coefficients larger than 0.995 and cosine similarities larger than 0.998, which indicates that the main CSI structure can be maintained with modest numerical accuracy. Degradation becomes nonlinear at lower precision (4 or 2 bits) as the coarse quantization intervals cannot express small but meaningful latent

Table 8

Robustness to latent quantization and feedback-link impairments.

Impairment	Condition	ΔNMSE (dB)	ρ	Cosine
Quantization	Float	0.000	0.996820	0.999680
	8 bit	2.690	0.996510	0.999410
	6 bit	5.530	0.995920	0.998870
	4 bit	15.330	0.985830	0.989270
	2 bit	29.900	0.831310	0.833450
Bit flips	$p = 10^{-4}$	4.160	0.995500	0.998460
	$p = 10^{-3}$	12.210	0.987020	0.990310
	$p = 10^{-2}$	22.060	0.911580	0.915360
Latent AWGN	30 dB	4.218	0.996256	0.999156
	20 dB	12.407	0.991299	0.994488
	10 dB	22.233	0.946800	0.950881
	0 dB	32.260	0.703943	0.702529

variations. Thus, the low-bit deployment can be aided by quantization-aware training instead of pure post-training quantization.

The feedback-bit experiment exhibits a similar tendency. At $p = 10^{-4}$, the correlation coefficient remains 0.99550, whereas increasing the error probability to 10^{-3} or 10^{-2} produces progressively larger losses. That is, it suggests that sporadic bit errors can be tolerated; however, concentrated random corruption interferes with compressed representation. Reliability can be enhanced through error-control coding or the differential protection of more sensitive codeword bits.

For latent AWGN, the model preserves high structural similarity at 30 dB and 20 dB, with $\rho = 0.996256$ and $\rho = 0.991299$, respectively. The performance also decreases significantly at 10 dB and 0 dB, where the noise energy is equivalent to the latent-signal energy. On average, WTSCNet performs well when subjected to mild feedback impairments; however, intense quantization, bit errors, or latent noise must be quantization-aware, coded, or noise-aware.

3.6 Deployment-oriented analysis

Since the encoder operates on the UE, its complexity and latency are particularly important. Table 9 presents the batch-one measurements. The central processing unit (CPU) latency was measured using a single thread, and the GPU latency was measured on an RTX 3090 with Compute Unified Device Architecture (CUDA) synchronization.

Table 9

Complexity and batch-one inference latency at $\eta = 1/4$.

Component	FLOPs	Parameters	Latency (ms)	
			CPU	RTX 3090
Encoder	2.269M	1.052M	1.064 ± 0.030	0.499 ± 0.026
Decoder	23.484M	1.073M	5.339 ± 0.136	2.643 ± 0.025
End-to-end	25.754M	2.124M	6.588 ± 0.101	3.154 ± 0.042

The encoder can only be used on 2.269M FLOPs, and the rest of the computation is allocated to the BS-side decoder. The 2.124M active parameters need about 8.10 MiB with a 32-bit floating point (FP32), 4.05 MiB with a 16-bit floating point (FP16), and 2.03 MiB with an 8-bit integer (INT8) representation. These findings justify asymmetric UE-BS deployment, but they cannot be tested directly on commercial mobile or embedded devices. Pruning, mixed-precision inference, and hardware-aware optimization are promising directions of deployment.

3.7 Spectral efficiency and downlink bit error rate

To evaluate whether CSI reconstruction accuracy translates into communication performance, we consider normalized single-user maximum-ratio transmission (MRT). The beamformer constructed from reconstructed CSI is

$$\hat{\mathbf{w}}_k = \frac{\hat{\mathbf{h}}_k}{\|\hat{\mathbf{h}}_k\|_2} \quad (14)$$

where $\hat{\mathbf{h}}_k$ is the reconstructed channel vector on the k -th subcarrier and $\|\cdot\|_2$ denotes the Euclidean (ℓ_2) norm. The normalized beamforming gain is

$$g_k = \frac{|\mathbf{h}_k^H \hat{\mathbf{w}}_k|^2}{\|\mathbf{h}_k\|_2^2} \quad (15)$$

where $\mathbf{h}_k$ denotes the true channel vector on the k -th subcarrier.

$$R = \frac{1}{N_c} \sum_{k=1}^{N_c} \log_2(1 + \gamma g_k) \quad (16)$$

$$P_{b,k} = Q(\sqrt{\gamma g_k}) = \frac{1}{2} \operatorname{erfc}\left(\sqrt{\frac{\gamma g_k}{2}}\right) \quad (17)$$

where R denotes the normalized SE, $P_{b,k}$ denotes the conditional uncoded gray-coded quadrature phase-shift keying (QPSK) bit error rate (BER) on the k -th subcarrier, γ is the downlink symbol signal-to-noise ratio (SNR), $Q(\cdot)$ is the Gaussian Q -function, and $\operatorname{erfc}(\cdot)$ is the complementary error function. Table 10 compares the reconstructed CSI with perfect CSI.

The mean normalized beamforming gain is 0.99394, and SE retention exceeds 99.5% over the evaluated SNR range. At 10 dB, the BER increases only from 7.8270×10^{-4} to 8.4888×10^{-4} . When the SNR is high, the relative BER difference can be seen as the perfect-CSI BER tends to approach zero despite the absolute WTSCNet BER being low. The findings apply only to uncoded single-user MRT under coherent equalization.

4 Discussion

The supplementary experiments explain the positive and negative aspects of WTSCNet. Module ablation proves that WTConv and CARBlock add to the reconstruction accuracy, whereas the wavelet-basis comparison indicates that the model does not depend much on the family of wavelets used in testing. Multi-seed outcomes show consistent correlation and cosine similarity, but mild NMSE training variability exists. Weak latent quantization, low bit errors, and high feedback-link SNRs retain the major CSI structure; however, strong impairments lead to obvious degradation.

The mixed-scenario experiment shows that joint indoor and outdoor samples can be reconstructed effectively, but the stringent cross-scenario findings show that the degradation is significant under zero-shot transfer. There are also other high-compression environments where WTSCNet is less competitive with TransNet+ as it has minimal explicit global dependency modeling. The next round of research will explore lightweight global modules, domain-resilient learning, quantization, and noise-sensitive training, channel coding, pruning, mixed-precision inference, and embedded UE deployment. The current assessment method does not incorporate controllable UE mobility, Doppler variation, or multi-cell interference. The COST2100 configuration considered in this study has static indoor and outdoor channel realizations but no explicit UE-velocity labels or configurable multi-cell interference topology. A proper analysis under such conditions would entail the use of a time-varying channel modeler with explicit assumptions on the structure of the cells, user scheduling, transmit power, coordination of interference, and Doppler spread. Thus, future work remains to be carried out where WTSCNet will be tested at different UE velocities, Doppler conditions, and inter-cell interference intensity. The combination of wavelet decomposition and spatial-channel processing can be extended to other spatial-frequency problems (image restoration, remote sensing, and intelligent-transportation perception) [19–21], in addition to CSI feedback.

5 Conclusions

WTSCNet is a network introduced in this study, which combines WT and attention as the means of improving feature representation with low computational complexity. The encoder uses WT to extract multi-resolution features, enabling better spatial and fine detail capture of features. In our decoder, we propose a synergistic module called SCSA that combines spatial and channel attention and adds it to CRBlock to form a CARBlock module. This combination successfully extracts and combines multi-scale, spatial, and channel-wise data, which is a more effective way of representing features and reconstruction accuracy. The experimental findings show that WTSCNet can be considered highly efficient in terms of reconstruction accuracy, outperforming the TransNet+ model at 1/4 compression rate and reducing computational complexity by 27.77%. We also tested reconstruction in heterogeneous channels, where we used a mixed dataset of COST2100 and compared WTSCNet with CLNet. It is evident that WTSCNet is still useful under joint availability of indoor and outdoor measurements.

Although WTSCNet has good performance at low compression ratios, it is required to perform more enhancements in order to be practically useful. Specifically, reconstruction with extremely high compression

Table 10

Normalized MRT SE and uncoded QPSK BER in the indoor scenario at $\eta = 1/4$.

SNR (dB)	SE (bit/s/Hz)		BER	
	Perfect-CSI	WTSCNet	Perfect-CSI	WTSCNet
0	1.0000	0.9956	1.5866×10^{-1}	1.5946×10^{-1}
5	2.0574	2.0507	3.7679×10^{-2}	3.8227×10^{-2}
10	3.4594	3.4514	7.8270×10^{-4}	8.4888×10^{-4}
15	5.0278	5.0193	9.3610×10^{-9}	2.7735×10^{-6}
20	6.6582	6.6495	7.6199×10^{-24}	4.1240×10^{-7}

is still problematic, and the stringent cross-scenario findings indicate that mixed-scenario training is not a surefire way to achieve zero-shot transfer to unobserved propagation conditions. The next direction of our research will focus on lightweight global modeling, domain-robust and noise-sensitive learning, model compression, practical implementation on the UE side, and testing in mobile conditions, Doppler shifts, and multi-cell interference.

Acknowledgments

This work was supported by the Tangshan Applied Basic Research Project under Grant No. 25130208B and the Key Project of North China University of Science and Technology under Grant No. ZD-JL-202512.

CRedit authorship contribution statement

Yong-Li An: Writing—review & editing, Supervision, Project administration, Funding acquisition. **Shuo-Yang Lu:** Writing—original draft, Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis. **Yi-Xiao Liu:** Writing—review & editing, Validation. **Ying-Chao Liu:** Writing—review & editing, Validation. **Qiang Kong:** Writing—review & editing, Methodology, Software, Validation, Formal analysis, Investigation, Visualization.

Declaration of competing interest

The authors declare no conflicts of interest.

References

- [1] P.-H. Kuo, H.T. Kung, P.-A. Ting, Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays, in: Proc. of the IEEE Wireless Communications and Networking Conf., Paris, France, 2012, pp. 492–497.
- [2] C.A. Metzler, A. Maleki, R.G. Baraniuk, From denoising to compressed sensing, *IEEE T. Inform. Theory* 62 (9) (2016) 5117–5144.
- [3] C.-B. Li, W.-T. Yin, Y. Zhang, User's guide for TVAL3: TV minimization by augmented Lagrangian and alternating direction algorithms, Version beta2.4 [Online]. Department of Computational and Applied Mathematics, Rice University, Houston, USA, Available, <https://usermanual.wiki/Document/UserGuide.2125550948.pdf>, 2010-11-11.
- [4] Y.-H. Zhang, Y.-F. Mu, Y. Liu, et al., Deep learning-based beamspace channel estimation in mmWave massive MIMO systems, *IEEE Wirel. Commun. Le.* 9 (12) (2020) 2212–2215.
- [5] H.-T. He, S. Jin, C.-K. Wen, et al., Model-driven deep learning for physical layer communications, *IEEE Wirel. Commun. Le.* 26 (5) (2019) 77–83.
- [6] J.-J. Guo, C.-K. Wen, S. Jin, et al., Overview of deep learning-based CSI feedback in massive MIMO systems, *IEEE T. Commun.* 70 (12) (2022) 8017–8045.
- [7] C.-K. Wen, W.-T. Shih, S. Jin, Deep learning for massive MIMO CSI feedback, *IEEE Wirel. Commun. Le.* 7 (5) (2018) 748–751.
- [8] J.-J. Guo, C.-K. Wen, S. Jin, et al., Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: design, simulation, and analysis, *IEEE T. Wirel. Commun.* 19 (4) (2020) 2827–2840.
- [9] Z.-L. Lu, J.-T. Wang, J. Song, Multi-resolution CSI feedback with deep learning in massive MIMO system, in: Proc. of the IEEE Intl. Conf. on Communications, Dublin, Ireland, 2020, pp. 1–6.
- [10] Q.-Y. Cai, C. Dong, K. Niu, Attention model for massive MIMO CSI compression feedback and recovery, in: Proc. of the IEEE Wireless Communications and Networking Conf., Marrakesh, Morocco, 2019, pp. 1–5.
- [11] S.-J. Ji, M. Li, CLNet: complex input lightweight neural network designed for massive MIMO CSI feedback, *IEEE Wirel. Commun. Le.* 10 (10) (2021) 2318–2322.
- [12] B.-H. Wang, Y.-L. Teng, V. Lau, et al., CCA-Net: a lightweight network using criss-cross attention for CSI feedback, *IEEE Commun. Le.* 27 (7) (2023) 1879–1883.
- [13] Y. Xu, M.-Q. Yuan, M.-O. Pun, Transformer empowered CSI feedback for massive MIMO systems, in: Proc. of the 30th Wireless and Optical Communications Conf., Taipei, China, 2021, pp. 157–161.
- [14] Y.-D. Cui, A.-H. Guo, C.-L. Song, TransNet: full attention network for CSI feedback in FDD massive MIMO system, *IEEE Wirel. Commun. Le.* 11 (5) (2022) 903–907.
- [15] Q. Chen, A.-H. Guo, Y.-D. Cui, Multi-head transformer architecture with higher dimensional feature representation for massive MIMO CSI feedback, *Applied Sciences* 14 (4) (2024) 1356.
- [16] Y.-L. An, S.-Y. Lu, H.-R. Cai, et al., A deep learning-based approach to lightweight CSI feedback, *Phys. Commun.-AMST* 68 (2025) 102538.
- [17] J.-X. Li, Y.-J. Wang, Z. Liu, et al., MetaFormer-based lightweight neural network for massive MIMO CSI feedback, *IEEE Wirel. Commun. Le.* 14 (2) (2025) 275–279.
- [18] X.-Y. Cen, C.-T. Lam, B.K. Ng, et al., When Transformer meets CSI feedback in mMIMO systems: a lightweight CsiMobileViT approach, in: Proc. of the IEEE 102nd Vehicular Technology Conf., Chengdu, China, 2025, pp. 1–6.

- [19] Y. Liu, G. Yang, T. Li, et al., Nighttime flare removal via wavelet-guided and gated-enhanced spatial-frequency fusion network, in: Proc. of the 40th AAAI Conf. on Artificial Intelligence, Singapore, 2026, pp. 7404–7412.
- [20] H. Zhou, Y.-L. Wang, Q. Zhang, et al., A dual-stage residual diffusion model with perceptual decoding for remote sensing image dehazing, *IEEE T. Geosci. Remote* 63 (2025) 4109312.
- [21] Y. Liu, X.-R. Wang, E.-P. Hu, et al., VNDHR: variational single nighttime image dehazing for enhancing visibility in intelligent transportation systems via hybrid regularization, *IEEE T. Intell. Transp.* 26 (7) (2025) 10189–10203.
- [22] Y.J. Kwon, A. Lee, H. Lee, CSI feedback compression based on deep learning using wavelet transform, in: Proc. of the 14th Intl. Conf. on Information and Communication Technology Convergence, Jeju Island, Republic of Korea, 2023, pp. 1091–1094.
- [23] Y.-Z. Si, H.-Y. Xu, X.-Z. Zhu, et al., SCSA: exploring the synergistic effects between spatial and channel attention, *Neurocomputing* 634 (2025) 129866.
- [24] L.-F. Liu, C. Oestges, J. Poutanen, et al., The COST 2100 MIMO channel model, *IEEE Wirel. Commun.* 19 (6) (2012) 92–99.



Yong-Li An received the Ph.D. degree from Beijing Jiaotong University, Beijing, China in 2015. She is currently a Professor with North China University of Science and Technology, Tangshan, China. She was a Visiting Scholar at Oakland University, Rochester, USA in 2011 and Griffith University, Brisbane, Australia in 2018. Her current research interests include key technologies of wireless communication physical layer, intelligent communication network, applications in the field of unmanned aerial vehicle (UAV) communication and authentication, and key technologies for UAV system channels.



Shuo-Yang Lu was born in Shijiazhuang, China in 2001. He obtained the M.S. degree from North China University of Science and Technology in 2026. His research interest mainly focuses on channel state information (CSI) feedback for massive multiple-input multiple-output (MIMO) systems.



Yi-Xiao Liu was born in Shijiazhuang, China in 2001. She received the B.S. degree from Qiqihar University, Qiqihar, China in 2019 and the M.S. degree from North China University of Science and Technology in 2026. Her main research interests include end-to-end communication systems and channel modeling.



Ying-Chao Liu received the Ph.D. degree from Yanshan University, Qinhuangdao, China in 2020. She is currently a Professor with North China University of Science and Technology. She was a Visiting Scholar at Northumbria University, Newcastle, UK in 2025. Her current research interest mainly focuses on intelligent optical fiber sensing technology integrating physical information neural networks.



Qiang Kong was born in Jining, China in 2002. He received the B.S. degree from Linyi University, Linyi, China in 2024. He is currently pursuing the M.S. degree at North China University of Science and Technology. His research interest mainly focuses on the radio frequency fingerprint security authentication of UAVs.