Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Deep unfolding based channel estimation for wideband terahertz near-field massive MIMO systems*

Jiabao GAO[1], Xiaoming CHEN‡[1], Geoffrey Ye LI[2]

[1]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*

[2]*Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BU, UK*

E-mail: gao_jiabao@zju.edu.cn; chen_xiaoming@zju.edu.cn; Geoffrey.Li@imperial.ac.uk

Received Nov. 8, 2023; Revision accepted Apr. 6, 2024; Crosschecked June 25, 2024; Published online Aug. 8, 2024

Abstract: The combination of terahertz and massive multiple-input multiple-output (MIMO) is promising for meeting the increasing data rate demand of future wireless communication systems thanks to the significant bandwidth and spatial degrees of freedom. However, unique channel features, such as the near-field beam split effect, make channel estimation particularly challenging in terahertz massive MIMO systems. On one hand, adopting the conventional angular domain transformation dictionary designed for low-frequency far-field channels will result in degraded channel sparsity and destroyed sparsity structure in the transformed domain. On the other hand, most existing compressive sensing based channel estimation algorithms cannot achieve high performance and low complexity simultaneously. To alleviate these issues, in this study, we first adopt frequency-dependent near-field dictionaries to maintain good channel sparsity and sparsity structure in the transformed domain under the near-field beam split effect. Then, a deep unfolding based wideband terahertz massive MIMO channel estimation algorithm is proposed. In each iteration of the approximate message passing-sparse Bayesian learning algorithm, the optimal update rule is learned by a deep neural network (DNN), whose architecture is customized to effectively exploit the inherent channel patterns. Furthermore, a mixed training method based on novel designs of the DNN architecture and the loss function is developed to effectively train data from different system configurations. Simulation results validate the superiority of the proposed algorithm in terms of performance, complexity, and robustness.

Key words: Terahertz; Massive MIMO; Channel estimation; Deep learning
https://doi.org/10.1631/FITEE.2300760 **CLC number:** TN92

1 Introduction

Terahertz (THz) massive multiple-input multiple-output (MIMO) is recognized as a promising technology in future wireless communication systems because the huge bandwidth and spatial degrees of freedom can support various emerging applications requiring high data rates (Wan et al., 2021; Hu et al., 2023). Nevertheless, this kind of appealing double gain heavily relies on the accuracy of available channel parameters, whose estimation is particularly challenging in THz massive MIMO systems for several reasons.

Compressive sensing (CS) algorithms are able to recover high-dimensional channels from low-dimensional received pilots with reduced overhead. However, the performance of CS-based channel estimators will be severely degraded if the special features of THz massive MIMO channels are not properly handled. On one hand, the receiver will easily fall in the near field of the electromagnetic wave sent by the transmitter, as illustrated in Fig. 1. The boundary to divide the near and far fields, defined as the Rayleigh distance, is proportional to the array

‡ Corresponding author

* Project supported by the National Key R&D Program of China (No. 2020YFB1805704)

ORCID: Xiaoming CHEN, https://orcid.org/0000-0002-1818-2135

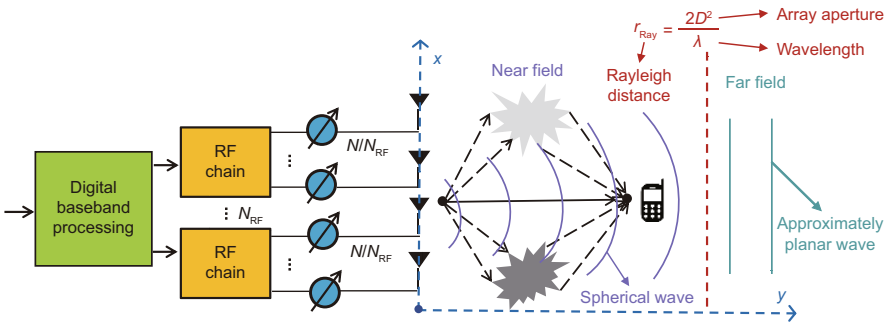

Fig. 1 Partially connected hybrid analog–digital massive MIMO system

aperture and inversely proportional to the wavelength (Cui and Dai, 2022). Within the range of the Rayleigh distance, the spherical wavefront must be considered, and simply using the approximately planar wavefront model will reduce the channel sparsity. In THz massive MIMO systems, the combination of short wavelength and large array aperture usually leads to Rayleigh distances of up to hundreds of meters, and thus cannot be ignored in most cellular systems. On the other hand, with both large bandwidth and array aperture, the beam split effect appears such that the equivalent angles corresponding to the same physical channel path are different at different subcarriers (Cui et al., 2023a). As a result, the widely exploited common sparsity structure among subchannels no longer holds. Last but not the least, the complexity of most existing high-performance CS algorithms becomes unbearable with a massive antennas array.

To deal with the near-field effect, Cui and Dai (2022) geometrically derived the near-field array response vector, which is dependent on not only the angle but also the distance. Then the near-field dictionary was constructed by multiple near-field array response vectors with different angle and distance grids. The classic CS algorithm with multiple measurement vectors (MMVs), simultaneous orthogonal matching pursuit (SOMP), was adopted for channel estimation exploiting the common sparsity structure. In Wei and Dai (2022), the hybrid-field scenario was considered, where the far- and near-field paths were estimated sequentially. For the mixed line-of-sight (LoS) and non-line-of-sight (NLoS) scenario, the LoS path component was first estimated in Lu and Dai (2023) in an off-grid manner, and then the NLoS path components were processed in an on-grid manner. To promote structured sparsity under the spatial non-stationary effect caused by the spherical wavefront and visibility region issues, the prior distribution of the orthogonal approximate message passing (OAMP) algorithm was tailored in Zhu et al. (2021). To compensate for the beam split effect, the dictionary was designed to be frequency-dependent in Elbir et al. (2023), so that the common sparsity structure can be maintained, whereas the bilinear pattern detection method proposed in Cui and Dai (2023) collects energy from all frequencies to determine the location of the grid with the highest total power in the angle–distance domain.

Deep learning (DL) has achieved great success in many wireless communication problems in the past few years (Qin et al., 2019). Recently, DL has been applied to near-field THz massive MIMO channel estimation. In Elbir et al. (2023), a black-box deep neural network (DNN) was used to predict wideband THz near-field channels based on the received pilot signals, whereas only the channel path key parameters including angles, distances, and gains were predicted in Chen et al. (2021). In Nayir et al. (2022), the coarse estimation of orthogonal matching pursuit (OMP) was refined using a denoising autoencoder to further improve estimation performance. To reduce the complexity, in Zhang et al. (2023), a smaller dictionary was learned together with parameters inserted into the iterative shrinkage and thresholding algorithm. To realize adaptive complexity and guarantee linear convergence, an efficient channel estimator was developed by Yu et al. (2022) based on a fixed-point DNN.

However, the above works either consider few practical narrowband scenarios or have limited performance. In addition, most existing DL-based methods are trained separately under different system configurations, which increases the training

and storage overhead and decreases the robustness. Therefore, estimators with high performance, low complexity, and strong robustness still need to be investigated for practical THz massive MIMO channels. In Gao et al. (2023a, 2023b), the superiority of deeply unfolding the advanced approximate message passing (AMP)-sparse Bayesian learning (SBL) algorithm was validated in wideband millimeter wave (mmWave) massive MIMO channel estimation. To extend this method to the THz band and enhance its robustness, we modify it in terms of dictionary design, network architecture, and training scheme in this study. Our main contributions are summarized as follows:

1. We use frequency-dependent near-field dictionaries to compensate for the near-field beam split effect of wideband THz massive MIMO channels, thus improving the performance of the proposed DL-based channel estimator.

2. We propose a deep unfolding based channel estimation algorithm, where AMP is used to lower the complexity of the SBL algorithm and the DL-based parameter update procedure in each iteration improves both the convergence speed and the performance.

3. We customize the DNN architecture to effectively exploit the inherent patterns of wideband THz massive MIMO channels in the angle–distance–frequency domain. Furthermore, the attention mechanism is applied to make the DNN adaptive to different system configurations by dynamically reweighting network features.

4. We creatively design a weighted normalized mean-squared error (NMSE) loss function to realize effective mixed training of data from different system configurations, so that we can obtain a single robust DNN that works well under various configurations.

Notations: We use italic, boldface lowercase, and boldface uppercase letters to denote scalar, vector, and matrix, respectively. $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $|\cdot|$, and $||\cdot||_{\mathrm{F}}$ denote the transpose, conjugate transpose, modulus, and Frobenius norm, respectively. $\mathrm{diag}(\cdot)$ converts a vector to a diagonal matrix. $\cdot^2$ and $./$ denote element-wise squaring and division, respectively. $\mathbf{1}_a$ and $\mathbf{0}_a$ denote the $(a,1)$-dimensional all-one vector and all-zero vector, respectively. $a:b:c$ denotes the arithmetic sequence vector starting from a and ending at c with common difference b. $\mathbb{C}^{x\times y}$ denotes the $x\times y$ complex space. $\mathcal{CN}(\mu,\sigma^2)$ denotes a circularly symmetric complex Gaussian (CSCG) random variable with mean μ and variance σ^2, whereas $\mathcal{CN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a CSCG random vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. $\mathcal{U}[a,b]$ denotes the uniform distribution between a and b.

2 System model

In this section, the massive MIMO system is first presented. Then, the wideband THz near-field channel model is introduced, after which the channel estimation process is formulated as a classic MMV-CS problem.

2.1 Massive MIMO system

As illustrated in Fig. 1, we consider a base station (BS) equipped with an N-antenna uniform linear array (ULA) and N_{RF} ($N_{\mathrm{RF}}\ll N$) radio frequency (RF) chains. To reduce the hardware overhead and power consumption, a partially connected hybrid analog–digital architecture is considered, where each RF chain is connected to N/N_{RF} antennas through N/N_{RF} low-cost one-bit phase shifters. We consider the frequency division duplex (FDD) system, where the BS transmits pilots to a single-antenna user (the extension to multiple multi-antenna users is straightforward because the pilots sent by the BS are broadcast to all user antennas) on K subcarriers for downlink channel estimation. Once the channels are estimated at the user, they will be fed back to the BS for effective beamforming, which further facilitates accurate downlink signal decoding. Without loss of generality, we assume that the baseband pilot fed to all the RF chains is always 1. Then the received signal at the user on the k^{th} subcarrier in the m^{th} time slot can be expressed as

$$y_m^k=\boldsymbol{w}_m^{\mathrm{T}}\boldsymbol{h}^k+n_m^k, \tag{1}$$

where $\boldsymbol{w}_m\in\mathbb{C}^{N\times 1}$, $\boldsymbol{h}^k\in\mathbb{C}^{N\times 1}$, and $n_m^k\sim\mathcal{CN}(0,\sigma^2)$ denote the phase shifts of the phase shifters at the BS, the channels from the BS to the user, and the additive noise with variance σ^2 at the user, respectively. Stacking the received signals of totally M time slots within which the channel is assumed to be constant, we further obtain

$$\boldsymbol{y}^k=\boldsymbol{W}\boldsymbol{h}^k+\boldsymbol{n}^k, \tag{2}$$

where $\boldsymbol{y}^k=[y_1^k,y_2^k,\cdots,y_M^k]^{\mathrm{T}}\in\mathbb{C}^{M\times 1}$, $\boldsymbol{n}^k=[n_1^k,n_2^k,\cdots,n_M^k]^{\mathrm{T}}\in\mathbb{C}^{M\times 1}$, and $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,$

$\boldsymbol{w}_M]^{\mathrm{T}} \in \mathbb{C}^{M\times N}$ whose elements are randomly selected from $\frac{1}{\sqrt{N}}\{+1,-1\}$ with an equal probability.

2.2 Wideband THz near-field channel model

Compared to the low-frequency counterpart, the wideband massive MIMO channel model in the THz band is much more complicated due to the near-field beam split effect, which makes the array response vector dependent on angles, distances, and the frequency of the subcarrier. Denote r as the distance from the center of the array at the BS to the scatter or the user and assume $r < r_{\mathrm{Ray}}$, where $r_{\mathrm{Ray}} = \frac{2D^2}{\lambda}$ denotes the Rayleigh distance (Cui and Dai, 2022) within which the near-field effect cannot be ignored. Furthermore, $D = (N-1)d$ denotes the array aperture with $d = \frac{\lambda}{2}$ denoting the antenna spacing, where $\lambda = \frac{c}{f_{\mathrm{c}}}$ denotes the carrier wavelength with c and f_{c} denoting the speed of light and the central frequency, respectively.

Straightforwardly, the near-field array response vector at frequency f can be expressed as

$$\begin{aligned}&\boldsymbol{a}_N(f,r,r^{(0)},\cdots,r^{(N-1)})\\&=\frac{1}{\sqrt{N}}\left[\mathrm{e}^{-\mathrm{j}2\pi\frac{c}{f}(r^{(0)}-r)},\cdots,\mathrm{e}^{-\mathrm{j}2\pi\frac{c}{f}(r^{(N-1)}-r)}\right]^{\mathrm{T}},\end{aligned}\tag{3}$$

where the distance from the n^{th} antenna of the array to the scatter or the user, $r^{(n)}$, can be calculated according to the geometry as $r^{(n)} = \sqrt{r^2-2r\delta_n d\theta+\delta_n^2 d}$ with $\theta \in [-1,1]$ denoting the sine of the angle of departure (AoD) at the BS and $\delta_n = \frac{2n-N+1}{2}, n = 0,1,\cdots,N-1$. Based on the Fresnel approximation, we further have $r^{(n)} \approx r-\delta_n d\theta+\frac{\delta_n^2 d^2(1-\theta^2)}{2r}$ (Cui and Dai, 2022). Therefore, Eq. (3) can be rewritten as

$$\begin{aligned}\boldsymbol{a}_N(\theta,r,f)=\frac{1}{\sqrt{N}}\Big[&\mathrm{e}^{-\mathrm{j}2\pi\frac{c}{f}\left(\frac{\delta_0^2 d^2(1-\theta^2)}{2r}-\delta_0 d\theta\right)},\\&\cdots,\mathrm{e}^{-\mathrm{j}2\pi\frac{c}{f}\left(\frac{\delta_{N-1}^2 d^2(1-\theta^2)}{2r}-\delta_{N-1} d\theta\right)}\Big]^{\mathrm{T}}.\end{aligned}\tag{4}$$

Using Eq. (4), the k^{th} downlink subchannel can be expressed as

$$\boldsymbol{h}^k=\sqrt{\frac{N}{N_{\mathrm{c}}N_{\mathrm{p}}}}\sum_{i=1}^{N_{\mathrm{c}}}\sum_{j=1}^{N_{\mathrm{p}}}\alpha_{i,j}\mathrm{e}^{-\mathrm{j}2\pi f_k\tau_{i,j}}\boldsymbol{a}_N(\theta_{i,j},r_{i,j},f_k),\tag{5}$$

where N_{c} and N_{p} denote the number of clusters and the number of subpaths in a cluster, respectively. In addition, $\alpha_{i,j},\tau_{i,j}$, and $\theta_{i,j}$ denote the path gain, delay, and sine of the AoD of the i^{th} subpath in the j^{th} cluster, respectively, whereas $r_{i,j}$ denotes the distance from the center of the BS array to the scatter or the user corresponding to the i^{th} subpath in the j^{th} cluster. Furthermore, we have $f_k = f_{\mathrm{c}}+(k-1-\frac{K-1}{2})\frac{f_{\mathrm{s}}}{K}$ with f_k and f_{s} denoting the frequency of the k^{th} subcarrier and the bandwidth, respectively, and $\theta_{i,j} = \sin\phi_{i,j}$ with $\phi_{i,j}$ denoting the physical AoD of the i^{th} subpath in the j^{th} cluster.

2.3 Formulation of the MMV-CS problem

To facilitate channel estimation, it is a common practice to first perform domain transformation with a proper dictionary in massive MIMO systems, so that channels in the transformed domain possess appealing properties such as sparsity and sparsity structures to help solve the underdetermined Eq. (2). To deal with the aforementioned near-field beam split effect in wideband THz channels, the dictionary should be designed as follows. On one hand, the atoms of the dictionary should naturally conform to the form of the near-field array response vector to enhance channel sparsity in the transformed domain (Cui and Dai, 2022). On the other hand, instead of using a common dictionary on all subcarriers, frequency-dependent dictionaries should be applied on different subcarriers to compensate for the beam split effect similar to the far-field case (Gao et al., 2023a). Specifically, the dictionary for the k^{th} subchannel is designed as

$$\boldsymbol{A}^k=[\boldsymbol{A}_0^k,\cdots,\boldsymbol{A}_{S-1}^k]\in\mathbb{C}^{N\times G},\tag{6}$$

where S denotes the number of sampled distance grids and the s^{th} sub-dictionary is further composed of Q near-field array response vectors as

$$\boldsymbol{A}_s^k=[\boldsymbol{a}_N(\theta_0,r_{s,0},f_k),\cdots,\boldsymbol{a}_N(\theta_{Q-1},r_{s,Q-1},f_k)],\tag{7}$$

where Q denotes the number of sampled angle grids. Therefore, the total number of grids in the transformed polar domain is $G = SQ$. According to Cui and Dai (2022), to minimize the maximum coherence between two arbitrary atoms to improve the performance of CS algorithms, the angles should be uniformly sampled while the distances should be non-uniformly sampled as

$$\theta_q=\frac{2q-Q+1}{Q},\ q=0,1,\cdots,Q-1,\tag{8}$$

$$r_{s,q}=\frac{1}{s}Z_{\triangle}(1-\theta_q^2),\ s=0,1,\cdots,S-1, \tag{9}$$

where $Z_{\triangle}=\frac{N^2d^2}{2\beta_{\triangle}^2 c/f_c}$ with $\beta_{\triangle}$ denoting the threshold that balances the coherence level and grid resolution, and S is set to the minimum integer that satisfies $\frac{1}{S}Z_{\triangle}<\rho_{\min}$ with $\rho_{\min}$ denoting the minimum allowable distance (Cui and Dai, 2022). Using the dictionary in Eq. (6), we can readily perform domain transformation as

$$\boldsymbol{h}^k\approx\boldsymbol{A}^k\boldsymbol{x}^k, \tag{10}$$

where $\boldsymbol{x}^k\in\mathbb{C}^{G\times 1}$ denotes the k^{th} sparse polar domain subchannel and $\approx$ is due to the quantization error caused by the finite grid resolution, which is usually very small in massive MIMO systems with dense grids (Gao et al., 2023b). As shown in Fig. 2, compared to using a common angular dictionary, using frequency-dependent polar dictionaries not only enhances the channel sparsity but also recovers the common sparsity structure among subchannels, thus effectively compensating for the near-field beam split effect.

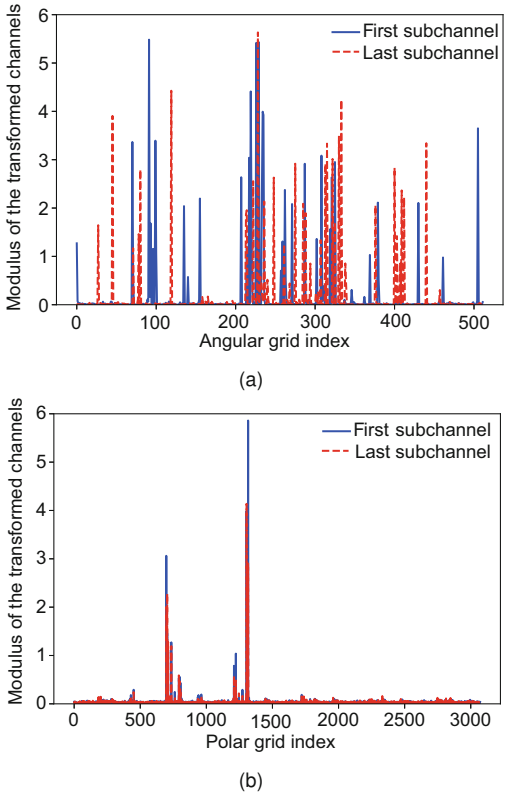

Fig. 2 Impact of dictionaries under the default simulation setting: (a) a common angular dictionary; (b) frequency-dependent polar dictionaries

Taking the $\approx$ in Eq. (10) by $=$ and substituting it into Eq. (2), we can obtain the following CS model:

$$\boldsymbol{y}^k=\boldsymbol{\Phi}^k\boldsymbol{x}^k+\boldsymbol{n}^k,\forall k, \tag{11}$$

where $\boldsymbol{\Phi}^k\triangleq\boldsymbol{W}\boldsymbol{A}^k$ denotes the measurement matrix on the k^{th} subcarrier. In summary, the goal of channel estimation is to accurately recover $\boldsymbol{x}^k$ based on $\boldsymbol{y}^k,\boldsymbol{\Phi}^k$, and σ^2. Because we have K measurements on K subcarriers and $\boldsymbol{x}^k$ have basically the same sparse supports, i.e., the common sparsity structure, it is a typical MMV-CS problem. After the sparse polar domain subchannels are estimated, the original subchannels can be readily reconstructed using Eq. (10).

3 Deep unfolding based channel estimator

In this section, the principles of the AMP-SBL algorithms are first introduced briefly. Then, the proposed deep unfolding based algorithm that enhances the capability of AMP-SBL using DL is described in detail, including its architecture, training scheme, and gains.

3.1 AMP-SBL

As one of the most powerful CS algorithms, SBL has high theoretical sparse recovery performance and is flexible for exploiting various sparsity structures (Srivastava et al., 2019). Omitting the subcarrier superscript, to recover a particular sparse polar domain subchannel $\boldsymbol{x}$, SBL first assumes that

$$\boldsymbol{x}\sim\mathcal{CN}(\boldsymbol{0}_G,\mathrm{diag}(\gamma_1,\gamma_2,\cdots,\gamma_G)), \tag{12}$$

where γ_g $(g=1,2,\cdots,G)$ denotes the Gaussian variance parameter of the g^{th} element of $\boldsymbol{x}$. Then, dozens of expectation–maximization (EM) based iterations are executed. In each iteration, the E-step computes the posterior mean and covariance of $\boldsymbol{x}$ given $\boldsymbol{y}$, based on which γ_g is updated by the M-step. After convergence, the eventual posterior mean of $\boldsymbol{x}$ is regarded as an estimate of it. The proven sparsity-promoting property of SBL ensures that $\boldsymbol{\gamma}$ will gradually become a sparse vector as the iteration progresses (Srivastava et al., 2019); i.e., most elements in $\boldsymbol{\gamma}$ will be close to zero after convergence and the indices of those non-zero elements indicate the existence of channel paths at the corresponding angle and distance.

Because the E-step in SBL involves the computation of matrix multiplications and inversions, its complexity is very high, especially in the THz massive MIMO channel estimation problem where the dimensions of matrices are quite large. To alleviate this issue, the AMP-SBL algorithm in Luo et al. (2021) uses an alternative realization of the E-step based on AMP. Because only matrix–vector multiplications are involved, the complexity is dramatically reduced compared to the original realization of the E-step. Furthermore, unitary preprocessing is applied to improve the robustness to general measurement matrices. The extension of AMP-SBL to the MMV case has been straightforwardly made in Luo et al. (2021), where a common Gaussian variance vector was adopted and updated according to information from all measurements to exploit the common sparsity structure. The detailed algorithm is demonstrated in Algorithm 1, where L denotes the number of iterations. Lines 1 and 2 represent unitary preprocessing, where singular value decomposition is executed on each measurement matrix; line 3 represents the initialization procedure; lines 5–11 represent the AMP-based E-step executed on all subcarriers in the l^{th} iteration; lines 12 and 13 represent the M-step in the l^{th} iteration. Notice that we adopt frequency-dependent measurement matrices here, causing some slight differences from the original algorithm proposed in Luo et al. (2021).

Algorithm 1 Approximate message passing-sparse Bayesian learning for multiple measurement vector-compressive sensing

Input: $\boldsymbol{y}^k, \boldsymbol{\Phi}^k, \forall k, \sigma^2, L$
Output: $\hat{\boldsymbol{x}}^k = \boldsymbol{\mu}_{\boldsymbol{x}^k}^L, \forall k$
1: $\forall k: \boldsymbol{\Phi}^k = \boldsymbol{U}^k \boldsymbol{\Sigma}^k \boldsymbol{V}^k$
2: $\forall k: \boldsymbol{r}^k = (\boldsymbol{U}^k)^{\mathrm{H}} \boldsymbol{y}^k,\ \boldsymbol{A}^k = (\boldsymbol{U}^k)^{\mathrm{H}} \boldsymbol{\Phi}^k$
3: $\boldsymbol{\mu}_{\boldsymbol{x}^k}^0 = \boldsymbol{0}_G, (\boldsymbol{s}^k)^0 = \boldsymbol{0}_Q, \forall k; \boldsymbol{\gamma}^0 = \boldsymbol{1}_G, \epsilon^0 = 0.001$
4: **for** $l = 1:1:L$ **do**
5: $\quad \forall k: \boldsymbol{\tau}_{\mathrm{p}}^k = |\boldsymbol{A}^k|^{\cdot 2} \boldsymbol{\tau}_{\boldsymbol{x}^k}^{l-1},\ \boldsymbol{p}^k = \boldsymbol{A}^k \boldsymbol{\mu}_{\boldsymbol{x}^k}^{l-1} - \boldsymbol{\tau}_{\mathrm{p}}^k (\boldsymbol{s}^k)^{l-1},$
6: $\qquad \boldsymbol{\tau}_{\mathrm{s}}^k = \boldsymbol{1}./(\boldsymbol{\tau}_{\mathrm{p}}^k + \sigma^2 \boldsymbol{1})$
7: $\quad \forall k: (\boldsymbol{s}^k)^l = \boldsymbol{\tau}_{\mathrm{s}}^k (\boldsymbol{r}^k - \boldsymbol{p}^k),$
8: $\qquad \boldsymbol{\tau}_{\mathrm{q}}^k = \boldsymbol{1}./\left(\left| (\boldsymbol{A}^k)^{\mathrm{H}} \right|^{\cdot 2} \boldsymbol{\tau}_{\mathrm{s}}^k \right),$
9: $\qquad \boldsymbol{q}^k = \boldsymbol{\mu}_{\boldsymbol{x}^k}^{l-1} + \boldsymbol{\tau}_{\mathrm{q}}^k \left((\boldsymbol{A}^k)^{\mathrm{H}} (\boldsymbol{s}^k)^l \right)$
10: $\quad \forall k: \boldsymbol{\mu}_{\boldsymbol{x}^k}^l = \boldsymbol{q}^k./(\boldsymbol{1} + \boldsymbol{\tau}_{\mathrm{q}}^k \boldsymbol{\gamma}^{l-1}),$
11: $\qquad \boldsymbol{\tau}_{\boldsymbol{x}^k}^l = \boldsymbol{\tau}_{\mathrm{q}}^k./(\boldsymbol{1} + \boldsymbol{\tau}_{\mathrm{q}}^k \boldsymbol{\gamma}^{l-1})$
12: $\quad \boldsymbol{\gamma}^l = \frac{2\epsilon^{l-1}+1}{\frac{1}{K}\sum_{k=1}^K (|\boldsymbol{\mu}_{\boldsymbol{x}^k}^l|^2 + \boldsymbol{\tau}_{\boldsymbol{x}^k}^l)}$
13: $\quad \epsilon^l = \frac{1}{2} \sqrt{\lg\left(\frac{1}{G} \sum_{g=1}^G \gamma_g^l \right) - \frac{1}{G} \sum_{g=1}^G \lg \gamma_g^l}$
14: **end for**

3.2 AMP-SBL unfolding

In practice, although unitary preprocessing improves the robustness to some extent, the AMP-SBL algorithm can still easily diverge under structured measurement matrices, such as those adopted in this study, making its low complexity meaningless. On the other hand, the M-step in the SBL algorithm is derived based on the assumption that the elements of $\boldsymbol{x}$ are independent of each other (Srivastava et al., 2019), which is not true in practical THz channels. With clusters and power leakage among grids (Gao et al., 2023b), the elements of $\boldsymbol{x}$ corresponding to close polar grids tend to have close modulus, which results in the block sparsity structure. In this case, the original M-step is far from optimal.

To overcome the shortcomings of AMP-SBL and obtain a channel estimation algorithm with good performance, low complexity, and strong robustness, we propose a deep unfolding based algorithm. Specifically, after the unitary preprocessing procedure, the EM-based iterations are unfolded into a large cascading DNN, each layer of which, named the AMP-SBL layer, corresponds to an iteration of the AMP-SBL algorithm. In each AMP-SBL layer, the AMP-based E-step remains unchanged, while the original simple M-step is replaced by a complicated function realized by a small DNN, whose architecture is carefully designed to effectively exploit the inherent channel patterns.

As illustrated in Fig. 3, in the l^{th} AMP-SBL layer, the AMP-based E-step is first executed on all K subcarriers, where $|\boldsymbol{\mu}_{\boldsymbol{x}^k}^l|^2, \boldsymbol{\tau}_{\boldsymbol{x}^k}^l$ are obtained based on the common $\boldsymbol{\gamma}^{l-1}, \sigma^2$ and the subcarrier-specific $\boldsymbol{A}^k, \boldsymbol{r}^k$ on the k^{th} subcarrier. After that, for each k, $|\boldsymbol{\mu}_{\boldsymbol{x}^k}^l|^2, \boldsymbol{\tau}_{\boldsymbol{x}^k}^l$ are reshaped into (S, Q)-dimensional matrices because the block sparsity structure exists in the two-dimensional (2D) polar domain, and are further stacked along the last expanded dimension to obtain the $(S, Q, 2)$-dimensional feature tensor. Then, the feature tensor is processed by a 2D convolutional (Conv) layer with 16 filters of size 5×5. Zero padding is executed to keep the dimensions unchanged after convolution. To exploit the common sparsity structure, K different $(S, Q, 16)$-dimensional tensors are then averaged to obtain a single $(S, Q, 16)$-dimensional tensor. Eventually, another 2D Conv layer with a single filter of size 5×5 is used to output the $(S, Q, 1)$-dimensional updated

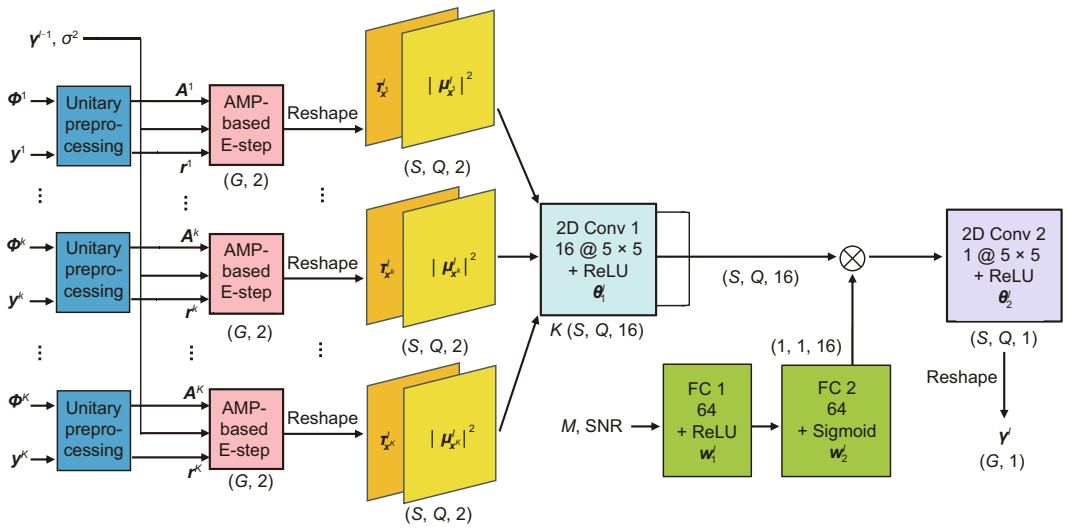

Fig. 3 Architecture of the l^{th} approximate message passing-sparse Bayesian learning layer

variance parameter tensor, which is reshaped back to a $(G, 1)$-dimensional vector $\boldsymbol{\gamma}^l$ to facilitate the next AMP-SBL layer's computation.

Most existing DNN-based channel estimators need to be trained separately under different system configurations, and the lack of robustness dramatically reduces their practical values. In contrast, the proposed algorithm naturally works with different system scales such as the number of time slots M, because the dimensions of the feature maps going through the Conv layers, i.e., S and G, are manually selected hyperparameters, and thus can be fixed. Furthermore, we add an attention module to the backbone network to enhance its adaptability to different configurations. Specifically, two fully connected (FC) layers are used to predict 16 weights according to configuration parameters, which are then multiplied to the feature maps before the second Conv layer in the backbone network along the last dimension. Notice that, although we consider only M and the signal-to-noise ratio (SNR) here as an example, many other configuration parameters can be included in practice. Through this dynamic feature re-weighting process, DNN can adapt to different configurations flexibly by changing the attention paid to different features (Gao et al., 2022). Overall, the function of the DNN-based M-step in the l^{th} AMP-SBL layer can be mathematically expressed as Eq. (13), where $f^l_{C_1}(\cdot;\boldsymbol{\theta}^l_1)$ and $f^l_{C_2}(\cdot;\boldsymbol{\theta}^l_2)$ denote the operations of the first and second Conv layers with weights $\boldsymbol{\theta}^l_1$ and $\boldsymbol{\theta}^l_2$, respectively, and $f^l_{D_1}(\cdot;\boldsymbol{w}^l_1)$ and $f^l_{D_2}(\cdot;\boldsymbol{w}^l_2)$ denote the operations of the first and second FC layers with weights $\boldsymbol{w}^l_1$ and $\boldsymbol{w}^l_2$, respectively. $g(x)=\max(0,x)$ denotes the ReLU activation function to enhance DNN's representation ability or guarantee the non-negativity of the updated Gaussian variances, and $\delta(x)=1/(1+\mathrm{e}^x)$ denotes the Sigmoid activation function to generate attention weights between 0 and 1.

$$\begin{aligned}\boldsymbol{\gamma}^l =& g\Big(f^l_{C_2}\Big(\delta\left(f^l_{D_2}\left(g\left(f^l_{D_1}\left(M,\mathrm{SNR};\boldsymbol{w}^l_1\right)\right);\boldsymbol{w}^l_2\right)\right)\\ &\cdot\frac{1}{K}\sum_{k=1}^{K} g\left(f^l_{C_1}\left(|\boldsymbol{\mu}^l_{\boldsymbol{x}^k}|^2,\boldsymbol{\tau}^l_{\boldsymbol{x}^k};\boldsymbol{\theta}^l_1\right)\right);\boldsymbol{\theta}^l_2\Big)\Big).\end{aligned} \tag{13}$$

3.3 Training scheme

The optimal $\boldsymbol{\theta}^l_1,\boldsymbol{\theta}^l_2,\boldsymbol{w}^l_1,\boldsymbol{w}^l_2,\forall l$ need to be obtained through training. We generate 8000, 1000, and 1000 channel samples with various system configurations as the training set, the validation set, and the testing set, respectively. Due to the stacking of multiple network layers and activation functions, the loss function is nonconvex and cannot be guaranteed to converge to the global optimum. To achieve good performance, the following training techniques are adopted. First of all, the Adam optimizer is adopted to exploit both first- and second-order momentums. Also, to avoid getting stuck in a bad local optimum, layer-wise training is adopted (Gao et al., 2023a), where shallower networks are trained first and deeper networks with newly added layers are trained on their basis. Last but not the least, the

weight of each newly added layer is initialized by the previous layer's weight, which reduces the loss oscillation and accelerates the convergence dramatically. When the performance does not increase with a new layer being added, the previous network without the newly added layer is the final algorithm. In simulation, the typical value of L is 10. In addition, the initial learning rate is set to 10^{-3}, and strategies including learning rate decay and early stopping are used in each training to improve the training speed and prevent overfitting. The batch size is set to 16 due to the memory limit of the graphics processing unit (GPU) used.

Because the channel estimation performance under different system configurations, such as M and SNR, varies a lot, configurations with better performance will be overwhelmed by configurations with worse performance during mixed training if we simply adopt NMSE as the loss function, which results in unbalanced training levels of different configurations (Gao et al., 2022). To deal with this issue, we propose the following weighted NMSE loss function:

$$\mathrm{Loss}(\boldsymbol{H},\hat{\boldsymbol{H}},M,\mathrm{SNR})=a(M,\mathrm{SNR})\frac{||\boldsymbol{H}-\hat{\boldsymbol{H}}||_{\mathrm{F}}^2}{||\boldsymbol{H}||_{\mathrm{F}}^2}, \tag{14}$$

where $\boldsymbol{H}\triangleq[\boldsymbol{h}^1,\boldsymbol{h}^2,\cdots,\boldsymbol{h}^K]$ denotes true subchannels, and $\hat{\boldsymbol{H}}$ denotes the subchannels outputted by the proposed algorithm. The weight corresponding to configuration parameters M and SNR, $a(M,\mathrm{SNR})$, is set to the ratio of the SBL algorithm's NMSE at a middle configuration point to that under M and SNR, so that smaller weights are assigned to the losses of harsher configurations, while larger weights are assigned to the losses of easier configurations. In this way, the loss levels of different configurations are balanced so that all configurations can achieve sufficient training, leading to effective mixed training.

3.4 Gain analysis

Compared to existing CS- and DL-based algorithms, the proposed algorithm has gains in the following aspects, which will also be validated by simulation results later:

1. First, because the approximation loss of the AMP-based E-step can be effectively compensated for by the properly trained DNN, the proposed algorithm can converge under structured measurement matrices where the original AMP-SBL algorithm diverges, thus making its low complexity meaningful.

2. The proposed algorithm is also promising for achieving better performance than SBL because DNN can exploit the actual channel distribution and learn a matched function of the M-step, which includes the original M-step as a special case. In addition, the solid foundation of the advanced SBL algorithm gives the proposed approach more potential to achieve great performance than other DL-based methods that are based on simple basic algorithms or are purely data-driven.

3. In terms of complexity, AMP simplifies the E-step per iteration, while the DNN-based M-step dramatically improves the convergence speed at the cost of moderate extra complexity. Therefore, the proposed approach has lower complexity than SBL and AMP-SBL, and its complexity is close to that of the simple SOMP (Cui and Dai, 2022). Specifically, the numbers of real floating operations (FLOPs) of SBL, AMP-SBL, the proposed algorithm, and SOMP are $16KM^2L_1G$, $20KML_2G$, $(20KM+800)LG$, and $8KML_3G$, respectively, where L_1, L_2, and L_3 denote the numbers of iterations of SBL, AMP-SBL, and SOMP, respectively.

4. Last but not the least, the numbers of required training samples and model parameters are quite small thanks to the model-driven nature. Also, the proposed mixed training method can dramatically reduce the DNN training and storage overhead at the user side in FDD systems without sacrificing performance compared to separate training. Specifically, an N_{config}-fold reduction can be achieved, where N_{config} denotes the number of possible configurations. Such a robust and consistent model is also simple to deploy and can reduce latency in fast-changing scenarios.

4 Simulation results

In this section, simulation results (for reproduction, the source code is available at https://github.com/EricGJB/Deep_Unfolding_terahertz_CE) are provided to validate the superiority of the proposed algorithm. The default settings are as follows unless specified: $N=256$, $K=32$, $f_{\mathrm{c}}=100$ GHz, and $f_{\mathrm{s}}=10$ GHz, so that $r_{\mathrm{Ray}}=97.5375$ m. Besides, $Q=512$, $\beta=1.2$, $\rho_{\min}=3$ m (Cui and Dai, 2022), so that $S=6$ and $G=3072$.

$N_c = 3$, $N_p = 10$, $\alpha_{i,j} \sim \mathcal{CN}(0,1)$, the angles and distances of subpaths within a cluster are assumed to obey Laplacian distributions whose means obey $\mathcal{U}[0°,\ 360°]$ and $\mathcal{U}[5\text{ m}, 30\text{ m}]$, respectively, while the standard deviations are set to 4° and 1 m, respectively (Cui et al., 2023b; Elbir et al., 2023). The SNR is defined as $1/\sigma^2$ and the middle configuration point is $M = 48, \text{SNR} = 10$ dB. All the performance points are obtained by averaging over 200 random data samples.

For benchmarks, we choose two on-grid CS algorithms, namely the SOMP algorithm (Cui and Dai, 2022; Elbir et al., 2023) and the MMV version of the original SBL algorithm, MSBL (Srivastava et al., 2019; Gao et al., 2023b), as well as their counterparts with different dictionaries and training schemes. To highlight the unique contributions of this study, off-grid CS algorithms and other DL-based algorithms are not compared, because the former usually have poor performance in cluster channels and the latter have shown obvious inferiority to the proposed approach in the mmWave band (Gao et al., 2023a, 2023b).

First of all, we would like to clarify that the hyperparameters in the proposed approach, including the network architecture, the optimizer, and the learning rate, are determined through simulations. Specifically, different hyperparameters are tried and the combination with the lowest validation loss is selected. For instance, Fig. 4 shows the impact of the optimizer on DNN training. As can be seen, Adam outperforms other optimizers in terms of convergence speed and performance. Due to limited space, the process of determining other hyperparameters is omitted here.

Fig. 5 shows the impact of dictionary under different M's, wherein different dictionaries are distinguished by line types. First of all, a larger M reasonably leads to better estimation performance of all algorithms, owing to more information gathered about the channels, while the pilot overhead grows as well. Although the polar dictionaries (PDs) lead to better performance than the angular dictionaries (ADs) in SOMP, owing to higher sparsity, the situation is reversed in MSBL due to larger coherence among atoms. However, in the proposed deep unfolding algorithm, the performance superiority of PD appears again, owing to the compensation effect of DNN, resulting in the lowest NMSE among all algorithms under various M's. This kind of performance superiority can also help save pilot overhead. For instance, to achieve a −9 dB NMSE, MSBL with AD requires $M = 64$, while AMP-SBL unfolding with PD requires only $M = 37$. Finally, notice that AMP-SBL is not visible in the figure because it has a high probability of divergence and thus terrible average NMSE under structured measurement matrices, which reflects the vital role of DNN in the proposed approach.

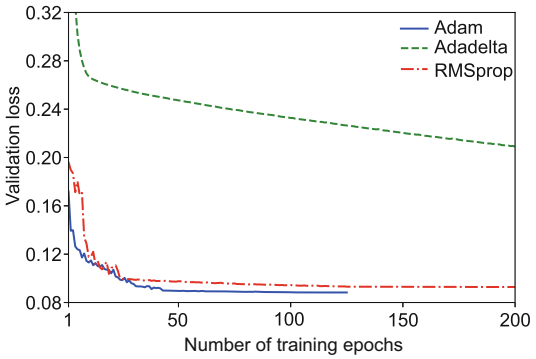

Fig. 4 Impact of the optimizer on DNN training (three AMP-SBL layers are unfolded)

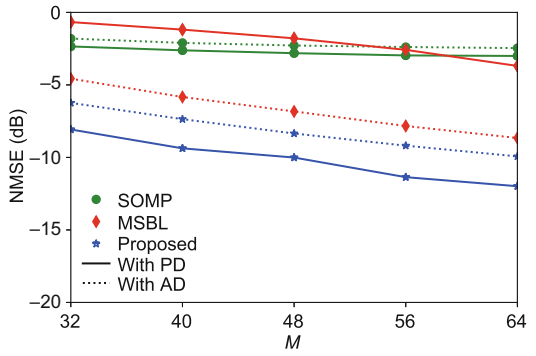

Fig. 5 NMSE versus M with different dictionaries

Furthermore, the behaviors of different algorithms under different levels of near-field beam split are investigated. Fig. 6 illustrates the impact of distance, wherein all subpath distances are set to a common value for convenient comparison, i.e., $r_{i,j} = r_c, \forall i, j$. Again, PD leads to better performance in SOMP and the proposed algorithm, while AD is better in MSBL. When r_c is small, the performance gaps between using PD and AD are large in both SOMP and the proposed algorithm, due to the relatively strong near-field effect. As r_c increases, the performance gaps gradually narrow and

eventually vanish when r_c is large enough. As for the impact of bandwidth, because the beam split effect is compensated for by the frequency-dependent measurement matrices and the common sparsity structure among subchannels always holds, the channel estimation performances of various algorithms rarely change with the bandwidth.

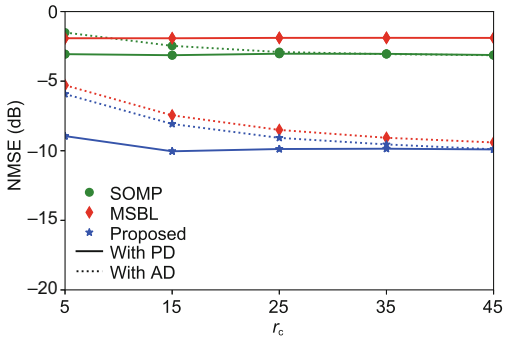

Fig. 6 NMSE versus r_c

Apart from better performance, the proposed approach has much lower complexity than MSBL with only about 1/37 FLOPs per iteration and 1/10 iteration number, as shown in Table 1. Owing to its DNN-based implementation, the proposed approach even runs faster than SOMP on the same CPU. In addition, with more iterations unfolded, more layers and more and larger kernels used in DNN in each iteration, the performance of the proposed approach gradually increases with stronger representation ability. However, the complexity increases at the same time. Finally, the performance curve reaches a plateau where further increasing the network scale leads to no performance improvement and even risk overfitting. In this study, hyperparameters at the turning point of the performance curve are selected to achieve the best performance with the least complexity. Nevertheless, it is totally feasible to adopt smaller network scales in practice to sacrifice performance to meet strict complexity requirements.

Table 1 Complexity comparison among different algorithms

Algorithm	FLOPs per iteration ($\times 10^9$)	Number of iterations	Overall FLOPs ($\times 10^9$)	Average running time (ms)
SOMP-PD	0.037	6	0.226	1.23
MSBL-AD	3.623	100	362.388	415.69
Proposed	0.097	10	0.978	0.89

Eventually, to verify the generality and robustness of the proposed algorithm, Fig. 7 shows the performances of different algorithms with different training schemes under various system configurations, including SNR and M (distinguished by line types). As can be seen from the figure, the performance superiority of the proposed algorithm over MSBL is maintained under various system configurations, validating its generality. Compared to separate training (ST) of a bunch of DNNs in different configurations, the proposed attention mechanism and the weighted NMSE loss function jointly realize effective mixed training (MT) of data from different configurations with very slight performance degradation at some configuration points, which is reflected by the close curves of ST and MT in the figure. Specifically, the original unnormalized NMSEs of the proposed approach at different configuration points differ by at most 28.5399 and will lead to unbalanced MT. Applying the weights calculated based on MSBL-AD's NMSEs, the NMSE difference factor decreases to only 2.0392, and the NMSEs of the proposed approch at all configuration points are basically at the same level around 0.1. Also, the amount of data of MT equals that at each configuration point in ST, so the total data amount of MT is much smaller than that of ST.

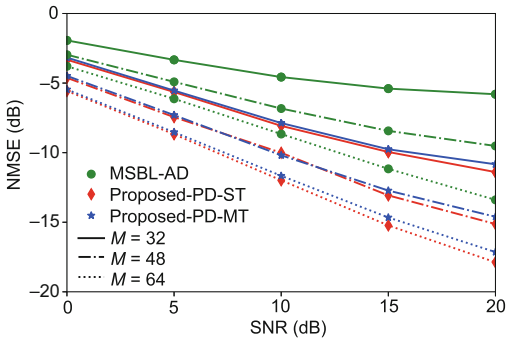

Fig. 7 NMSE under different system configurations

5 Conclusions and future work

In this study, we propose a deep unfolding based algorithm for wideband THz near-field massive MIMO channel estimation. We first compensate for the near-field beam split effect using frequency-dependent near-field domain transformation dictionaries. Then, we enhance the capability of the AMP-SBL algorithm using a DNN to learn the optimal

parameter update rule in each of its iterations. The DNN architecture is customized to exploit inherent channel patterns. Finally, we propose an effective MT method based on novel DNN architecture and loss function designs to obtain a single robust network that can work under various configurations. Simulation results demonstrate the good performance, low complexity, and strong robustness of the proposed algorithm. In the future, we will apply the proposed algorithm to more wireless communication problems and prove the convergence property to improve its universality and reliability.

Contributors

Jiabao GAO performed the simulations and drafted the paper. Xiaoming CHEN and Geoffrey Ye LI helped organize the paper. Xiaoming CHEN and Geoffrey Ye LI revised and finalized the paper.

Conflict of interest

Xiaoming CHEN is a corresponding expert of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Chen YH, Yan LF, Han C, 2021. Hybrid spherical- and planar-wave modeling and DCNN-powered estimation of terahertz ultra-massive MIMO channels. *IEEE Trans Commun*, 69(10):7063-7076. https://doi.org/10.1109/TCOMM.2021.3098696

Cui MY, Dai LL, 2022. Channel estimation for extremely large-scale MIMO: far-field or near-field? *IEEE Trans Commun*, 70(4):2663-2677. https://doi.org/10.1109/TCOMM.2022.3146400

Cui MY, Dai LL, 2023. Near-field wideband channel estimation for extremely large-scale MIMO. *Sci China Inform Sci*, 66(7):172303. https://doi.org/10.1007/s11432-022-3654-y

Cui MY, Dai LL, Wang ZC, et al., 2023a. Near-field rainbow: wideband beam training for XL-MIMO? *IEEE Trans Wirel Commun*, 22(6):3899-3912. https://doi.org/10.1109/TWC.2022.3222198

Cui MY, Tan JB, Dai LL, 2023b. Wideband hybrid precoding for THz massive MIMO with angular spread. *Sci Sin Inform*, 53(4):772-786. https://doi.org/10.1360/SSI-2022-0137

Elbir AM, Shi W, Papazafeiropoulos AK, et al., 2023. Near-field terahertz communications: model-based and model-free channel estimation. *IEEE Access*, 11:36409-36420. https://doi.org/10.1109/ACCESS.2023.3266297

Gao JB, Hu M, Zhong CJ, et al., 2022. An attention-aided deep learning framework for massive MIMO channel estimation. *IEEE Trans Wirel Commun*, 21(3):1823-1835. https://doi.org/10.1109/TWC.2021.3107452

Gao JB, Zhong CJ, Li GY, 2023a. AMP-SBL unfolding for wideband mmWave massive MIMO channel estimation. IEEE Int Conf on Communications Workshops, p.60-65. https://doi.org/10.1109/ICCWorkshops57953.2023.10283596

Gao JB, Zhong CJ, Li GY, et al., 2023b. Deep learning-based channel estimation for wideband hybrid mmWave massive MIMO. *IEEE Trans Commun*, 71(6):3679-3693. https://doi.org/10.1109/TCOMM.2023.3258484

Hu XL, Liu CX, Peng MG, et al., 2023. IRS-based integrated location sensing and communication for mmWave SIMO systems. *IEEE Trans Wirel Commun*, 22(6):4132-4145. https://doi.org/10.1109/TWC.2022.3223428

Lu Y, Dai LL, 2023. Near-field channel estimation in mixed LoS/NLoS environments for extremely large-scale MIMO systems. *IEEE Trans Commun*, 71(6):3694-3707. https://doi.org/10.1109/TCOMM.2023.3260242

Luo M, Guo QH, Jin M, et al., 2021. Unitary approximate message passing for sparse Bayesian learning. *IEEE Trans Signal Process*, 69:6023-6039. https://doi.org/10.1109/TSP.2021.3114985

Nayir H, Karakoca E, Görçin A, et al., 2022. Hybrid-field channel estimation for massive MIMO systems based on OMP cascaded convolutional autoencoder. Proc IEEE 96$^{\text{th}}$ Vehicular Technology Conf, p.1-6. https://doi.org/10.1109/VTC2022-Fall57202.2022.10013010

Qin ZJ, Ye H, Li GY, et al., 2019. Deep learning in physical layer communications. *IEEE Wirel Commun*, 26(2):93-99. https://doi.org/10.1109/MWC.2019.1800601

Srivastava S, Mishra A, Rajoriya A, et al., 2019. Quasi-static and time-selective channel estimation for block-sparse millimeter wave hybrid MIMO systems: sparse Bayesian learning (SBL) based approaches. *IEEE Trans Signal Process*, 67(5):1251-1266. https://doi.org/10.1109/TSP.2018.2890058

Wan ZW, Gao Z, Gao FF, et al., 2021. Terahertz massive MIMO with holographic reconfigurable intelligent surfaces. *IEEE Trans Commun*, 69(7):4732-4750. https://doi.org/10.1109/TCOMM.2021.3064949

Wei XH, Dai LL, 2022. Channel estimation for extremely large-scale massive MIMO: far-field, near-field, or hybrid-field? *IEEE Commun Lett*, 26(1):177-181. https://doi.org/10.1109/LCOMM.2021.3124927

Yu WT, Shen YF, He HT, et al., 2022. Hybrid far- and near-field channel estimation for THz ultra-massive MIMO via fixed point networks. IEEE Global Communications Conf, p.5384-5389. https://doi.org/10.1109/GLOBECOM48099.2022.10001564

Zhang XY, Wang ZN, Zhang HY, et al., 2023. Near-field channel estimation for extremely large-scale array communications: a model-based deep learning approach. *IEEE Commun Lett*, 27(4):1155-1159. https://doi.org/10.1109/LCOMM.2023.3245084

Zhu YF, Guo HY, Lau VKN, 2021. Bayesian channel estimation in multi-user massive MIMO with extremely large antenna array. *IEEE Trans Signal Process*, 69:5463-5478. https://doi.org/10.1109/TSP.2021.3114999