Hindawi
IET Signal Processing
Volume 2024, Article ID 6666395, 25 pages
https://doi.org/10.1049/2024/6666395

Research Article

A DOA Estimation Method Based on an Improved Transformer Model for Uniform Linear Arrays with Low SNR

Wei Wang,[1] **Lang Zhou**,[2] **Kun Ye**,[1] **Haixin Sun**,[1] **and Shaohua Hong**[1]

[1] *School of Informatics, Xiamen University, Xiamen, China*
[2] *School of Electronic Science and Engineering, Xiamen University, Xiamen, China*

Correspondence should be addressed to Haixin Sun; hxsun@xmu.edu.cn

Received 24 July 2023; Revised 14 November 2023; Accepted 18 January 2024; Published 17 February 2024

Academic Editor: Jung-Chieh Chen

In this paper, the Star-Transformer model is improved to obtain more accurate direction of arrivals (DOA) estimation of underwater sonar uniform linear array (ULA) under low signal-to-noise ratio (SNR) conditions. The ideal real covariance matrix is divided into three channels: real part channel, imaginary part channel, and phase channel to obtain more input features. In training, the real covariance matrix is used under different SNRs. In testing, the covariance matrix of samples in the real environment is used as input. The on-grid form is used to estimate the DOA of multiple signal sources, which is modelled as a multilabel classification problem. The results show that the model can be effective and can still have a good DOA estimation performance under the conditions of trained and untrained SNRs, different snapshots, signal power mismatch, different separation angles, signal correlation, and so on. It shows that the model has excellent robustness.

1. Introduction

DOA estimation technology has been well developed over the past few decades. It has been widely used in civil and military applications such as radar, sonar, wireless communications, underwater target monitoring, enemy ship detection, etc. In 1969, Capon proposed the minimum variance distortionless response (MVDR) method [1], which still has good resolution and performance under the influence of noise. Schmidt then proposed the multiple signal classification (MUSIC) algorithm [2], which greatly improved the performance of DOA estimation. Through eigen-value decomposition (EVD) and singular-value decomposition (SVD), noise subspaces of the signal correlation matrix are used to separate signals from different sources to achieve good DOA estimation. The excellent effect of this algorithm makes subspace methods widely used and many MUSIC-based algorithms appear [3, 4]. Subsequently, root-multiple signal classification (R-MUSIC) was proposed [5], which made the MUSIC-based algorithm break the grid limitation and obtain DOA estimation by solving the roots of polynomials, making it the most classical MUSIC-based algorithm. In 1985, the estimation of signal parameters via rotational invariance techniques (ESPRIT) proposed by Roy and Kailath [6] was applied for DOA estimation. Unlike the MUSIC method, which is solved by the peak of the pseudospectrum, it obtains the off-grid DOA by SVD, constructing matrix pairs and evaluating them by EVD. Many variations of this algorithm have followed [7, 8]. The MUSIC-based algorithm and the ESPRIT-based algorithm still have excellent performance when sparsity can be obtained for nested matrices and mutual prime matrices [9–11].

Compressed sensing (CS) [12] has been used in the field of direction of arrival (DOA) estimation for almost two decades and has shown significant progress while still evolving [13]. This type of method usually comes in three forms: on-grid, off-grid, and grid-less [14]. The on-grid and off-grid methods are more balanced, and their performance is inferior to the grid-less methods, but they have less computational complexity. The grid-less algorithm ensures the accuracy of the DOA estimation by sacrificing computational complexity. Due to the high complexity of the latter, it is sometimes not possible to detect it in real time. The DOA estimation of CS usually needs to solve the sparse minimisation problem based on ℓ_0 pseudonorm or (and) ℓ_1 norm, and the DOA estimation with high accuracy can be obtained through several iterations. The $\ell_{2,1}$-SVD algorithm greatly reduces the computation and can converge quickly [15, 16]. It reduces the computation by using dimensionality reduction so that it does not have to iterate over all the received signals in the snapshots. In the article [17], the spatial filter is constructed adaptively using the received signals, the weighted matrix is obtained using the spatial spectrum, and the weighted $\ell_{2,1}$ norm penalty is used for DOA estimation. It can achieve excellent performance with fewer snapshots. CS-based methods are sensitive to SNR or (and) the number of snapshots because they usually need to adjust parameters related to SNR or (and) the number of snapshots. It is difficult to obtain good results under the condition of SNR variation. In the literature [18], a block sparse Bayesian learning (BSBL) algorithm is proposed for the case where the mutual coupling is unknown, which does not require a separate computation of the mutual coupling and thus does not lead to aperture loss.

In recent years, numerous studies have been conducted regarding antijamming and adaptation to underwater environments. Literature [19] implemented a correction for localisation estimations in uncorrected arrays. Literature [20] used path attenuation and phase difference between sensor components to achieve three-dimensional estimation without requiring knowledge of the random unknown path loss exponent of an arbitrary nonfree space model. Literature [21] conducted a computational reduction of the MUSIC algorithm. Literature [22] proposed estimation methods adapted to correlated and uncorrelated sources. Literature [23] used the generalised search pattern (GPS) for generalised noise reconstruction, which provides good estimation in all kinds of noise.

The latest DOA estimation method is based on deep learning (DL). Compared to the traditional optimisation algorithm, it has the following advantages: (a) no continuous optimisation is required after training the network; (b) the network usually uses multiplication and addition, so it runs faster and reduces a lot of processing time than the optimisation algorithm; (c) the network has some robustness and can still have a good effect under the condition of fewer snapshots and lower SNR; and (d) all network parameters are obtained during the training process, eliminating the need to identify a small number of parameters that have a significant impact on the overall outcome, as required by other optimisation algorithms. In [24], the counter-diagonal element of the phase difference matrix in the frequency domain of the received data is used as the input to the convolutional neural networks (CNN) to estimate the parameters of different numbers of mixed sources and obtain the off-grid DOA estimation. In [25], the CNN network is used to obtain vectors that can construct the Toeplitz matrix, and R-MUSIC or Vandermonde decomposition is used to obtain DOA estimates. It is an off-grid algorithm and can accurately determine the number of sources. In [26], the circular array DOA estimation based on CNN is proposed, but the network has a large number of layers. In [27], using the real and imaginary parts of the upper right non-diagonal part of the covariance matrix as input to the CNN showed good generalization and accuracy, but its resolution was 3°. In [28], it is proposed that the network is both simple and has good single-source detection performance for different networks with large or small array elements. In [29], multilayer CNN is used for source number and DOA estimation, which has a good DOA estimation effect, and there is no need to adjust SNR and the number of snapshots, which makes the on-grid model more universal. Literature [30] describes a method for estimating the DOA in millimetre-wave multiple-input multiple-output (MIMO) systems using CNN without prior knowledge of the number of multipaths. Literature [31] introduces 2D DOA estimation algorithms for dilated convolutional networks that can adapt to low elevation angles.

The contribution of the work in this thesis is the first use of the transformer-based method to obtain good DOA estimates at low SNR. In the case of low SNR, the sample covariance matrix (SCM) is often very different from the real array manifold matrix due to the loud noise, which makes it difficult to estimate. To reduce the impact of the above problems, this paper has done the following work: (a) improve the Star-Transformer model in [32] and use the improved model to extract features, and the transformer-based method has better feature extraction effect than the CNN method [33]; (b) use multichannel data so that the network can extract features better; and (c) use dropout layers to improve model generalisation better and avoid over fitting. The main contributions of this paper are summarised as follows:

(1) In this paper, multichannel enhanced Star-Transformer data are used for DOA estimation for the first time. The multichannel consists of three channels: real part channel, imaginary part channel, and phase channel. The first two parts are the real and imaginary parts of the complex terms of the covariance matrix, whereas the phase channel is the phase of it and the value is $[-\pi, \pi]$. The covariance matrix of different SNRs under ideal conditions is used as training, so that it can adapt to the SCM under different SNRs and different snapshotsand perform excellent multisource DOA estimation. The measured DOA is discretized by a grid. The DOA estimation task is modelled as a multilabel classification task. The proposed model has a faster processing and running speed compared to the traditional and CNN algorithms. With the same amount of data as training samples, the proposed model can train faster on the data and requires fewer epochs to converge. For the CNN algorithm, which requires 200 epochs to converge, the proposed model requires only 50 epochs to reach convergence, and the average processing time per epoch is also smaller than that of the CNN method.

(2) Training with the ideal covariance matrix, although the results may be slightly worse than with SCM, the training sample and training time are reduced. The results show that the model can not only adapt to the situation of low SNR but also has a good effect in the case of high SNR and signal mismatch.

Our results show that (a) the proposed method performs better than other algorithms at low SNR, (b) this method can perform DOA estimation well when the source separation is

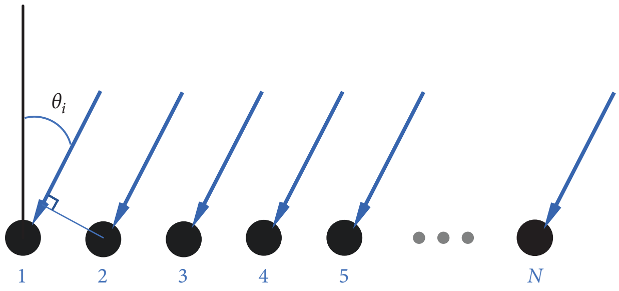


Figure 1: ULA signal incidence model.

small, and (c) the algorithm is robust under different SNRs, different number of snapshots, and signal mismatches.

The rest of the paper is structured as follows: the system model is presented in Section 2, the proposed network model based on the Star-Transformer is presented in Section 3, the training methods and related parameters are explained in Section 4, the simulation experiments are carried out in Section 5, and algorithm and possible future research directions are summarised in Section 6.

Notation: The following notation is used in this paper: $\mathscr{X}$ represents a set and $|\mathscr{X}|$ represents the cardinality of $\mathscr{X}$. $\mathbf{A}$ is a matrix, $\mathbf{a}$ is a vector, and a is a scalar. $(\mathbf{X}_{i,j})$ is denoted as the ith and jth element of $\mathbf{X}$. $\mathbf{x}(i)$ is represented as the ith element of the vector $\mathbf{x}$. The imaginary number has the unit j ($j^2 = -1$). The conjugate transpose of the matrix is $(\cdot)^H$, its conjugate is $(\cdot)^*$, and its transpose is $(\cdot)^T$. Lowercase italics are used to represent functions, e.g., $f(\cdot)$. $\mathbb{E}[\cdot]$ is expressed as the expectation operator. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ are represented as the real and imaginary parts of a complex object, respectively. $\angle\{\cdot\}$ is the phase angle of a complex object. $\mathbf{I}_N$ is the identity matrix of dimension $N \times N$. The distribution of white circular symmetric complex Gaussian noise with mean μ and covariance σ is expressed as $\mathscr{CN}(\mu, \sigma)$.

2. System Model

In this paper, the underwater scene is mainly considered, and the receiving array used is ULA. The array interval is $d = \frac{\lambda}{2}$, where $\lambda = \frac{c}{f}$ is the wavelength at the frequency f with the speed of sound c. The linear array model is shown in Figure 1.

According to the model mentioned in Figure 1, in the case of narrow band, the received signal model of K signal sources of N-element sensor array is generally expressed as

$$\begin{aligned} y(t) &= \sum_{k=1}^{K} a(\theta_k) s_k(t) + n(t) \\ &= \boldsymbol{A}(\theta)\boldsymbol{s}(t) + n(t), t = 1, \ldots, T, \end{aligned} \tag{1}$$

where $\boldsymbol{A}(\theta) = [a(\theta_1), a(\theta_2), \ldots, a(\theta_K)]$ is the $N \times K$ array manifold matrix, $\theta = [\theta_1, \theta_2, \ldots, \theta_K]^T$ is the unknown source direction vector, T denotes the total number of snapshots, $\boldsymbol{s} = [s_1, s_2, \ldots, s_K]^T$ represents the transmission signal, and $n(t)$ is the additive noise received at sampling index t. The columns of the array manifold matrix are as follows:

$$a(\theta_k) = \left[1, e^{j\frac{2\pi d}{\lambda}\sin(\theta_k)}, \ldots, e^{j\frac{2\pi d}{\lambda}(N-1)\sin(\theta_k)}\right]^T. \tag{2}$$

In the narrow-band far-field case, the following classical assumptions are usually made:

(i) The DOA of different signal sources is different

(ii) Each signal is randomly generated (Gaussian signalling) and uncorrelated [34], so there is a diagonal source covariance matrix $\mathbf{R}_s$, satisfying $\mathbf{R}_s = \mathbb{E}[\mathbf{s}(t)\mathbf{s}(t)^H] = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2)$, where σ_i^2 represents the power of the ith signal source

(iii) The additive noise is an independent but uniformly distributed additive white Gaussian noise, satisfying $n(t) \sim \mathscr{CN}(\mathbf{0}, \sigma_e^2 I_N)$, and is independent of the signal source

(iv) There is no time correlation between any two snapshots.

To obtain a correct estimate of the DOA of the unknown signal source (θ) from the received measurement data $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(T)]$, combined with the above assumptions, the covariance matrix of the received signal should satisfy the following equation:

$$\mathbf{R}_y = \mathbb{E}[\mathbf{y}(t)\mathbf{y}(t)^H] = \mathbf{A}(\theta)\mathbf{R}_s\mathbf{A}^H(\theta) + \sigma_e^2 I_N. \tag{3}$$

where $\mathbf{R}_y$ can be estimated from $K \leq N-1$ sources. In practice, however, $\mathbf{R}_y$ cannot be measured directly and can only be estimated by receiving signals in the case of finite snapshots:

$$\widetilde{\mathbf{R}}_y = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)\mathbf{y}(t)^H, \tag{4}$$

where $\tilde{\mathbf{R}}_y$ is an unbiased estimate of $\mathbf{R}_y$. $\tilde{\mathbf{R}}_y$ is very close to $\mathbf{R}_y$ for the maximum number of snapshots.

Training data of our model can be obtained after $\mathbf{R}_y$ processing, and the generated model can still estimate K sources θ from the received $\tilde{\mathbf{R}}_y$. If the hypothesis changes, such as the correlation between signal sources, the model may also have some robustness, but the greater the change, the greater the deviation, which is the same as the classical DOA estimation algorithm.

3. DOA Estimation Network Based on Improved Star-Transformer

This section describes the details and models of the multilabel DOA classification task. Section 3.1 explains the data processing and label format. Section 3.2 presents the Star-Transformer model and its functionality. Finally, Section 3.3 introduces the loss function used.

3.1. Data Processing and the Form of the Labels. This paper focuses on scenarios where the DOA fall within the range of [−60°, 60°], with a network resolution of 1°. When the DOA

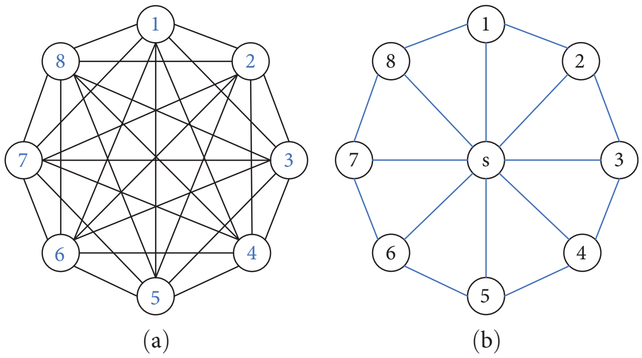


FIGURE 2: The different connection relationships between (a) the transformer proposed by Vaswani et al. [33] and (b) the Star-Transformer proposed by Guo et al. [32].

is between −90° and 90°, the linear array exhibits phase ambiguity in the regions of −90° to −60° and 60° to 90°. Although significant efforts were made to reduce errors in the two mentioned intervals, occasional deviations still resulted in a large root-mean-square error (RMSE). Therefore, the grid used is $\vartheta = \{-60^\circ, -59^\circ, \ldots, -1^\circ, 0^\circ, 1^\circ, \ldots, 59^\circ, 60^\circ\}$, with $|\vartheta| = 121$. For each SNR, K angles are selected from ϑ and the ideal covariance matrix is generated according to Equation (3) as training data. Because the neural network method cannot handle complex numbers and the ideal covariance matrix deviates largely from the actual covariance matrix at low SNRs, the use of a phase layer can mitigate the error to some extent. The data must be processed again to convert it into the three-channel data $\mathbf{X} \in \mathbb{C}^{N\times N\times 3}$, comprising the real part, imaginary part, and phase channels, i.e., $\mathbf{X}_{:,:,1} = \mathrm{Re}\{\mathbf{R}_y\}$, $\mathbf{X}_{:,:,2} = \mathrm{Im}\{\mathbf{R}_y\}$ and $\mathbf{X}_{:,:,3} = \angle\{\mathbf{R}_y\}$. Therefore, the input to the network is $\mathscr{X} = \{\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(D)}\}$, where D is the total number of data.

To generate label vectors, select K positions and assign them a value of 1, while assigning a value of 0 to the remaining positions. For example, there are two sources $\{-60^\circ, -59^\circ\}$, the resulting label vector is $\mathbf{z} = [1,\ 1,\ 0,\ \ldots,\ 0]^T$, and z is a 121×1 binary vector.

Finally, according to the data and label vector mentioned above, the training data set is obtained as follows:

$$\mathscr{D} = \left\{ \left(\mathbf{X}_{(1)}, \mathbf{z}_{(1)}\right), \left(\mathbf{X}_{(2)}, \mathbf{z}_{(2)}\right), \ldots, \left(\mathbf{X}_{(D)}, \mathbf{z}_{(D)}\right) \right\}, \tag{5}$$

which is of size D. Note that the ideal covariance in Equation (3) used in the training phase is known, whereas the covariance used in the verification and test phases is not known. The input covariance of the test phase should be the covariance in Equation (4), which is then transformed by the same processing into a three-channel data before being processed by the network.

3.2. The Proposed Improved Star-Transformer Model. The proposed model is a slight improvement on the Star-Transformer of [32]. The different connection relationships between the transformer proposed by Vaswani et al. [33] and the Star-Transformer proposed by Guo et al. [32] are shown in Figure 2. Each data node in transformer connects to each other in pairs, whereas each data node in Star-Transformer connects only to adjacent data nodes and virtual nodes in between. Transformer has good performance on large data samples, but its performance is sometimes hampered on small and medium samples. The use of ideal covariance requires that the model be trained in an environment with a small number of samples. Therefore, Star-Transformer, which can accommodate small and medium samples, give better results.

Before entering the Star-Transformer, the covariance matrix data, which are processed into three channels, must be dimensionally modified to make it more suitable for input. After the dimensional changes, the number of rows of data is padsize and the number of columns of data is d_{model}. Transformer-based methods usually require the inclusion of an embedding layer to make the model more sensitive to relative position [35]. However, the embedding layer did not improve performance, so it was discarded.

Unlike other multihead attention (MHA) networks in transformers, the proposed model simplifies the MHA network. For most transformers, the input data must be transformed into $\mathbf{Q}$ (query), $\mathbf{K}$ (key), and $\mathbf{V}$ (value) matrices by the weight matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$. The resulting $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrices are then run through the scaled dot-product attention function to obtain the attention score:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{6}$$

where

$$\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) = \frac{e^{\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}}}{\sum e^{\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}}}, \tag{7}$$

and d_k is the dimension of $\mathbf{K}$. In MHA, its value is equal to $\frac{d_{\mathrm{model}}}{m}$, where m is the number of heads.

Star-Transformer makes some improvements to Equation (6) by changing the attention score to the following formula:

$$\mathrm{STA}(\mathbf{q}, \mathbf{H}) = \mathrm{Attention}\left(\mathbf{q}\mathbf{W}^Q, \mathbf{H}\mathbf{W}^K, \mathbf{H}\mathbf{W}^V\right), \tag{8}$$

where $\mathbf{q}$ and $\mathbf{H}$ are the inputs of the MHA layer. The matrix $\mathbf{Q}$ can be obtained from $\mathbf{q}$, whereas the matrices $\mathbf{K}$ and $\mathbf{V}$ can be obtained from $\mathbf{H}$.

The proposed model further simplifies Equation (8) to obtain faster training and testing speed and simplifies the model to adapt to smaller and simpler data. The three weight matrices in Equation (8), which must be trained to be derived, are replaced by a constant known matrix **P** of size $d_{\mathrm{model}} \times \frac{d_{\mathrm{model}}}{m}$. For the change matrix $\mathbf{P}_l$ of the lth head, there is

$$\mathbf{P}_l(i,j) = \begin{cases} 1, \; i = (l-1) \cdot \dfrac{d_{\mathrm{model}}}{m} + j, \\ 0, \; \text{other} \end{cases} \tag{9}$$

Therefore, the formula (8) is transformed into

$$\text{Simple A}(\mathbf{q}, \mathbf{H}) = \text{Attention}(\mathbf{qP}, \mathbf{HP}, \mathbf{HP}). \tag{10}$$

Thus, the output of the improved MHA layer with m heads is represented as follows:

$$\text{Multi A}(\mathbf{q}, \mathbf{H}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_m)\mathbf{W}^O, \tag{11}$$

where $\mathbf{W}^O$ represents the weight matrix to be trained, $\text{Concat}(\cdot)$ indicates the concatenation operation, and

$$\begin{aligned} \text{head}_l &= \text{Simple A}_l(\mathbf{q}, \mathbf{H}) \\ &= \text{Attention}(\mathbf{qP}_l, \mathbf{HP}_l, \mathbf{HP}_l), \; l \in [1, m]. \end{aligned} \tag{12}$$

The improved MHA module is followed by the dropout layer, which sets the weight to 0 with a probability of 10%, and these weights cannot be obtained by training. This makes the network learn the data rather than memorise it, and can effectively prevent overfitting. The enhanced MHA and dropout layers together form the MHA-like layer. The ReLU function is needed after each MHA-like layer as an activation function.

Layer normalization (LN) [36] immediately after the MHA-like layer and activation function to achieve a more effective and generic model. The LN can normalise the data based on the sample mean and standard deviation.

The result obtained after the above changes is the same as the dimension of the input matrix. In general, it is necessary to add the data obtained from the LN layer to the input from the MHA-like layer to achieve the residual network effect and obtain better results. The proposed model uses this link in the calculation of the satellite nodes but not in the calculation of the relay nodes.

Similar to Star-Transformer in [32], the data **E** obtained by the dimensional changes are cloned in $\mathbf{H}_o$, and the pooling of **E** is performed to obtain the variable **s**. The anterior and posterior positions of $\mathbf{H}_o$ are labelled as $\mathbf{H}_n$ and $\mathbf{H}_l$, respectively. The five matrices $\mathbf{H}_n$, $\mathbf{H}_o$, $\mathbf{H}_l$, **E**, and **s** are integrated into a whole **C** as the **H** matrix of the MHA-like layer, whereas the data **E** themselves are the **q** matrix. Then, the satellite nodes are obtained by the activation function ReLU and the LN layer. The output of the LN layer **h** is added to **E** to obtain the updated $\mathbf{H}_1$. The $\mathbf{H}_1$ and **E** are combined into **M** as the **H** matrix of the MHA-like layer, whereas **E** is the **q** matrix. Like the satellite nodes, the data of the relay nodes are obtained through the MHA-like layer and the subsequent operations. The output is obtained after updating both nodes together many times. It is found that the proposed model has the best effect when the dimension is transformed into row vector, namely padsize = 1, and the number of cyclic updates is 1. In this way, the **H** input of the satellite nodes is only related to **E**. However, to prevent the output of Softmax from being a matrix of single elements $\{1\}$, five variables are still reserved for better results.

After the above operation, the data must pass through the position wise feed forward (PFF) network. The PFF network used by the model is represented by three fully connected (FC) layers. After the first two FC layers, GELU is used as the activation function, instead of the usual ReLU. The output of the PFF network is obtained by adding the output of the third FC layer to the input of the PFF network. The neurons in these three FC layers are 512, 256, and d_{model}. The output of the FC layer is shown below:

$$\mathbf{c}^{[n]} = \mathbf{W}^{[n]}\mathbf{c}^{[n-1]} + \mathbf{b}_{FC}^{[n]}, \tag{13}$$

where $\mathbf{c}^{[n]}$ and $\mathbf{c}^{[n-1]}$ are the output of the nth FC layer and the $(n-1)$th FC layer, respectively (the output of the previous layer is the input of the current layer), whereas $\mathbf{W}^{[n]}$ and $\mathbf{b}_{FC}^{[n]}$ are the weight matrix and the deviation between the output and input of the FC layer, respectively.

To achieve the desired result, the data output from the PFF layer is transformed into one dimension through the FC layers FC1, FC2, and FC3. The FC layers consist of 512, 256, and 121 neurons, respectively.

Finally, after the Sigmoid layer, whose function is $s(x) = e^x/(e^x + 1)$, the output vector is obtained. Each element of this vector corresponds to the probability of each of the discriminated angles, and its value is $[0, 1]$. The output is expressed as

$$\widehat{\mathbf{p}}_{(i)} = \begin{pmatrix} \widehat{\mathbf{p}}_1 \\ \widehat{\mathbf{p}}_2 \\ \vdots \\ \widehat{\mathbf{p}}_{121} \end{pmatrix}, \tag{14}$$

where $\widehat{\mathbf{p}}_{(i)}$ is the ith sample estimate and $\widehat{\mathbf{p}}_i$ is the possibility of the ith resolution angle.

The proposed network structure is shown in Figure 3.

3.3. Loss Function. The proposed model uses the supervised offline training method to build a multilabel classification model. Binary cross-entropy loss is used as the loss function, i.e.,

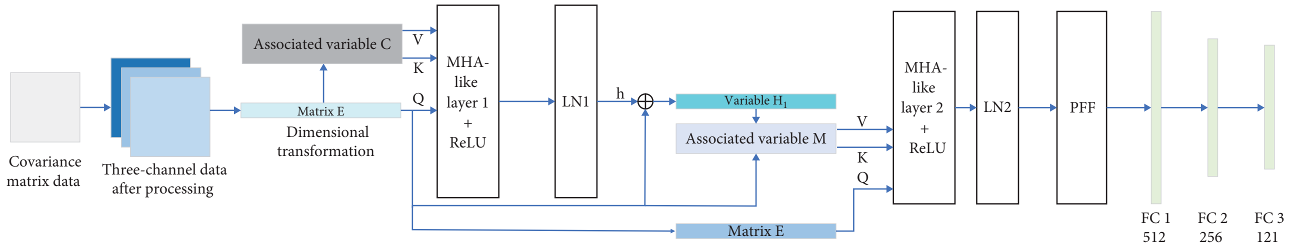


FIGURE 3: This paper proposes a model structure based on the Star-Transformer. After dimensional changes, the data must pass through the MHA-like layer with ReLU as the activation function and the LN layer with different association variables to obtain satellite nodes and relay nodes, respectively. After passing through the PFF layer, the output of the relay nodes must pass through the flatten layer to obtain one-dimensional output and pass through the FC layer. Sigmoid is used in the last FC layer to obtain the resolution angle probability. The MHA-like layer uses the dropout layer. The PFF layer and the updating of the satellite nodes use the residual structure.

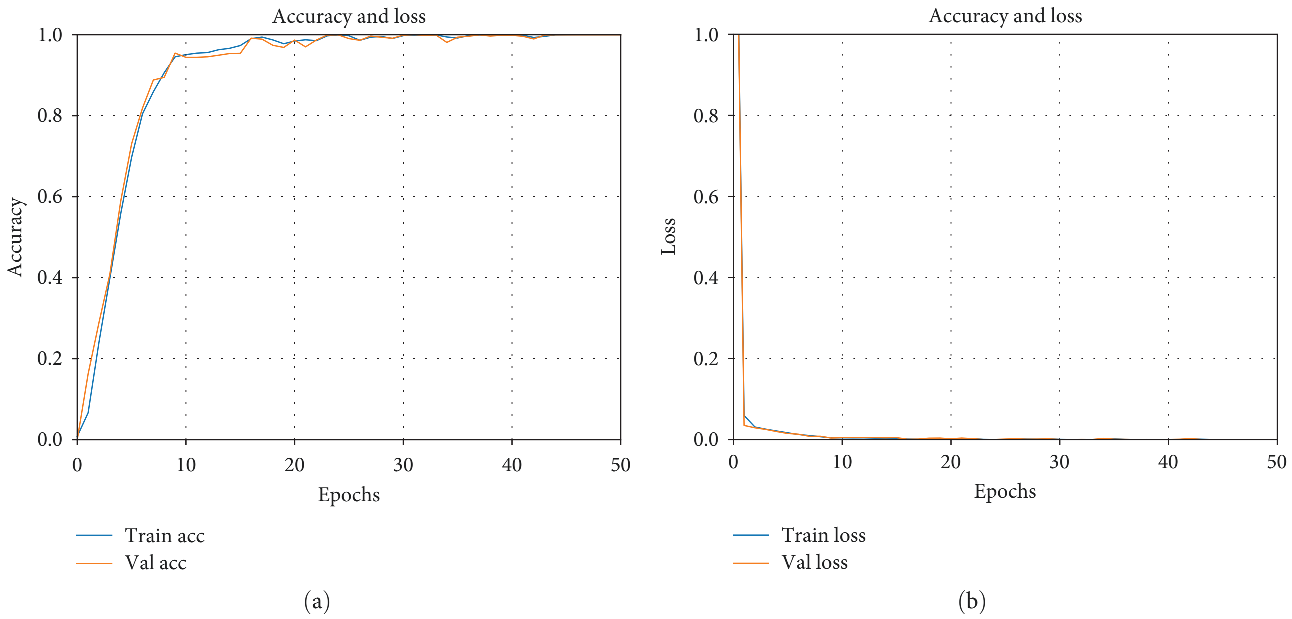


FIGURE 4: Accuracy and loss of training and validation for each epoch when the number of sources is $K = 2$.

$$\text{loss} = \arg\max_{\mathcal{G}} \frac{1}{D} \sum_{i=1}^{D} L\left(\widehat{\mathbf{p}}_{(i)}; \mathbf{z}_{(i)}\right), \tag{15}$$

where $\mathcal{G}$ represents all trainable parameters, and

$$\begin{aligned} L\left(\widehat{\mathbf{p}}_{(i)}; \mathbf{z}_{(i)}\right) &= -\frac{1}{|\vartheta|} \sum_{n=1}^{|\vartheta|} \Big[\mathbf{z}_{(i)}(n) \log\left(\widehat{\mathbf{p}}_{(i)}(n)\right) \\ &\quad + \left(1 - \mathbf{z}_{(i)}(n)\right) \log\left(1 - \widehat{\mathbf{p}}_{(i)}(n)\right)\Big]. \end{aligned} \tag{16}$$

4. Training Methods and Associated Parameters

The ideal covariance matrix under different SNRs is used as the training sample. It has the following advantages: (a) the sample size is significantly reduced, (b) there is no need to train the model for different numbers of snapshots or SNRs, (c) training time is greatly reduced, and (d) the generated model performs similar to the model trained on a single number of snapshots and SNR. In this paper, the case of source number $K = 2$ is considered. For $|\vartheta| = 121$, a total of $121 \times 120/2 = 7,260$ samples could be generated for each SNR. For the training set, five SNR values are chosen to match the low SNR environment: $\{-20, -15, -10, 5, 0\}$ db. Therefore, according to Equation (3), there is a total of $7,260 \times 5 = 36,300$ data as training samples. It was determined through experimentation that the model obtained from the training is also well suited for high SNR environments and varying SNRs. The model is well suited for real-world environments due to its ability to maintain performance as SNR and number of snapshots change.

For the above training sample, the data were randomly divided into a training set (90% of the sample) and a verification set (10% of the sample). Figure 4 shows the accuracy and loss of the training and verification under each epoch with source number $K = 2$. The verification set here is just to check if the model can learn data and if it is overfitting. The verification set is very different from the data in the real environment: (a) the test data in the real environment are estimated using the SCM as an input, not the ideal covariance matrix, which is unknown in real situations and (b) the tested angle pairs are generally not located on the grid, unlike the validation set, which only contains data from grid angles.

To update and optimise of the proposed model, Adam [37] was used with an initial learning rate of 0.001 and $\beta_1 = 0.9$, $\beta_2 = 0.999$. Once the loss stopped decreasing for six consecutive epochs, the learning rate decreased by a factor of 0.8. The batch size was set to 64, and the network was trained for 50 epochs. The model was trained using PyTorch in PyCharm. The operating system used is Windows, running on an Intel(R) Core(TM) i7-7700HQ CPU at 2.80 GHz and NVIDIA GeForce GTX 1050Ti GPU.

5. Simulation Experiment and Results

Extensive experiments have been performed to evaluate the DOA estimation performance of the proposed model under different conditions. All experiments except for those in Section 5.8 assume a known number of sources and $K = 2$. In the training and testing processes, the ULA of array elements $N = 16$ is used, and the distance between two adjacent array elements of it is half wavelength d. Each MHA-like layer has 590,592 parameters to be trained, each PFF layer has 722,432 parameters, the FC1 layer has 393,728 parameters, the FC2 layer has 131,328 parameters, the FC3 layer has 31,097 parameters, for a total of nearly 2.5 million parameters to be trained for the experimental model performed, and the SNR is defined in [38]:

$$\mathrm{SNR} = 10\log_{10} \frac{\min\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\}}{\sigma_e^2}. \tag{17}$$

At the test stage, the covariance matrix $\tilde{\mathbf{R}}_y$ of the measurement data is obtained according to Equation (4), and these data must also be converted to three channel.

5.1. The Algorithm Used for the Comparison. The following is a list of the algorithms used in this paper compared to the proposed algorithm:

(a) MUSIC in [2].
(b) TLS-ESPRIT in [6].
(c) $\ell_{2,1}$-SVD in [16]
(d) CNN suggested in [29].
(e) the simple CNN network proposed in [28], called CNNsimple.
(f) the MIMO CNN network proposed in [30], called CNNMIMO.

Methods (a) and (b) are classical algorithms in DOA estimation, method (c) is a compressed sensing method, and methods (d), (e), and (f) are neural network methods. For the on-grid methods (a), (c), (d), (e), and (f), the grid resolution is set to 1°, which is the same as the proposed model. All DL methods use the same amount of training volume. ESPRIT uses maximum array overlap and the total least square (TLS) method to implement the algorithm [39].

Table 1: Training time for DL algorithm.

DL algorithm	Training time (s)
CNN	7930.075
CNNsimple	728.761
CNNMIMO	1607.372
The proposed method	432.112

The Cramér-Rao lower bound (CRLB) used is described in [34].

In the test, the form of the output is similar to Formula (14), which is

$$\widehat{\widehat{\mathbf{p}}}_{(i)} = f(\widetilde{\mathbf{X}}) = \begin{pmatrix} \widehat{\widehat{\mathbf{p}}}_1 \\ \widehat{\widehat{\mathbf{p}}}_2 \\ \vdots \\ \widehat{\widehat{\mathbf{p}}}_{121} \end{pmatrix}. \tag{18}$$

The output of the test $\widehat{\widehat{\mathbf{p}}}_i$ represents the probability that the DOA of the real environment covariance matrix is the ith angle. In the case where the number of sources is known, the K angles with the highest probability are the DOA estimated by the model.

For all comparison and proposed algorithms, the performance was evaluated using the RMSE under Monte Carlo experiment, which is defined as follows:

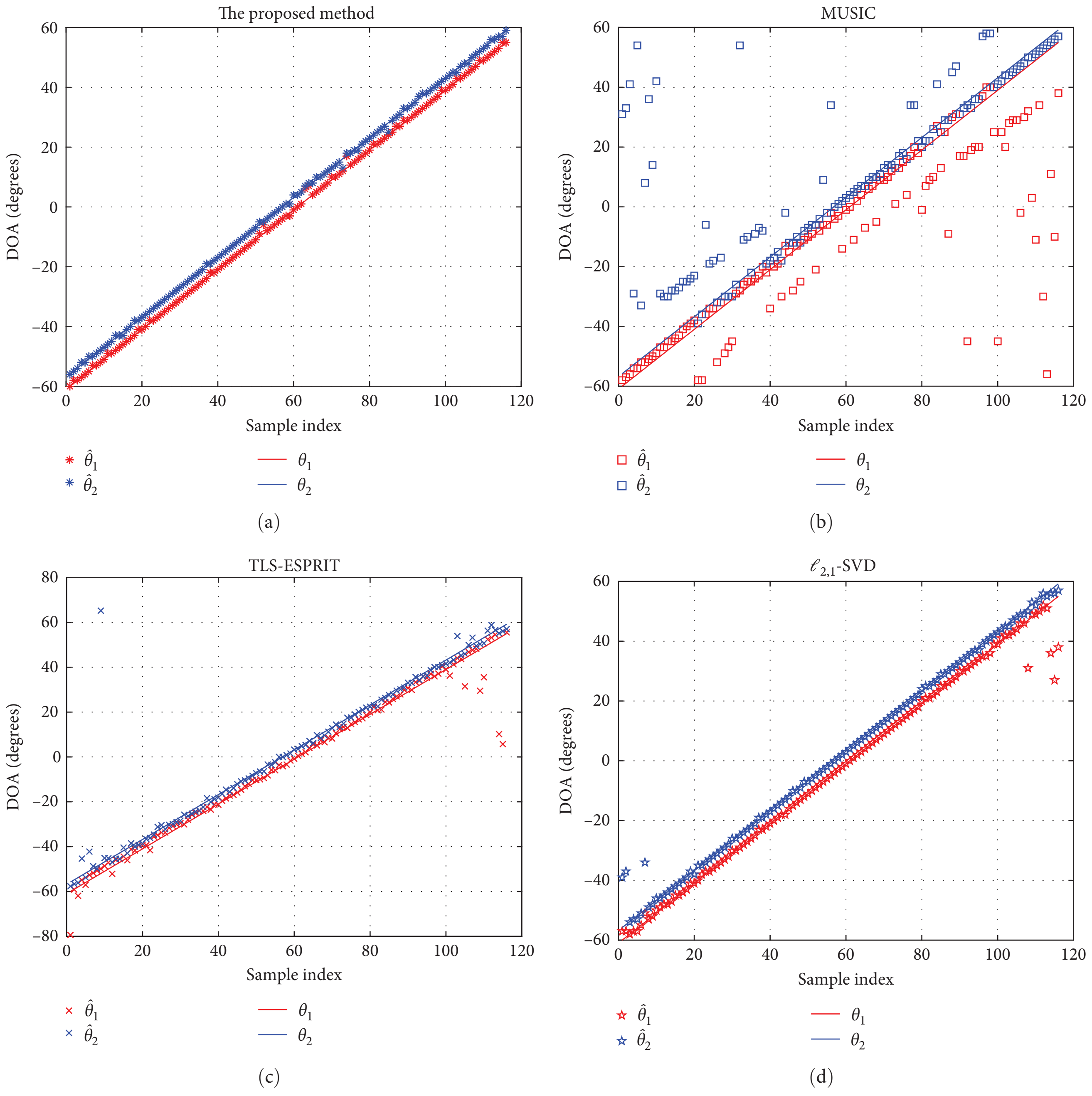
$$\mathrm{RMSE} = \sqrt{\frac{1}{D_{\text{test}}K} \sum_{r=1}^{D_{\text{test}}} \sum_{k=1}^{K} \left(\theta_k^{(r)} - \widehat{\theta}_k^{(r)}\right)^2}, \tag{19}$$

where $[\theta_1^{(r)}, \ldots, \theta_K^{(r)}]^T$ are the real DOA angles of the rth experimental sample, $[\widehat{\theta}_1^{(r)}, \ldots, \widehat{\theta}_K^{(r)}]^T$ are the DOA angles estimated by the rth experimental sample, and D_{test} represents the number of repetitions of a test experiment. Both the actual DOA and the estimated DOA are listed from smallest to largest, i.e., $\theta_1^{(r)} \le \theta_2^{(r)} \le \ldots \theta_K^{(r)}$ and $\widehat{\theta}_1^{(r)} \le \widehat{\theta}_2^{(r)} \le \ldots \widehat{\theta}_K^{(r)}$. If a large number of different cases are used and the experiment is not repeated, Formula (19) can be used with D_{test} representing the number of cases.

For DL algorithms, the training time for each algorithm is shown in Table 1. The proposed method has minimal training time.

5.2. DOA Estimation Performance and Error. Two experiments were conducted to evaluate the performance and error of DOA estimation. They reflect the cases of having a source on the grid and both sources being off-grid, respectively. In both experiments, it is assumed that the power of the two signals is the same, and there is no SNR mismatch problem.

The initial experiments focus on estimating the DOA at low SNR. The SNR of the two signals is −10 dB with 500 snapshots, and the angular spacing used is $\Delta\theta = 4.2°$. The first signal, θ_1, is a signal on the grid starting at −60° in steps of 1° and ending at 55°. For each θ_1, $\theta_2 = \theta_1 + \Delta\theta$. Only one angle is placed on the grid, so that the angle pairs are not



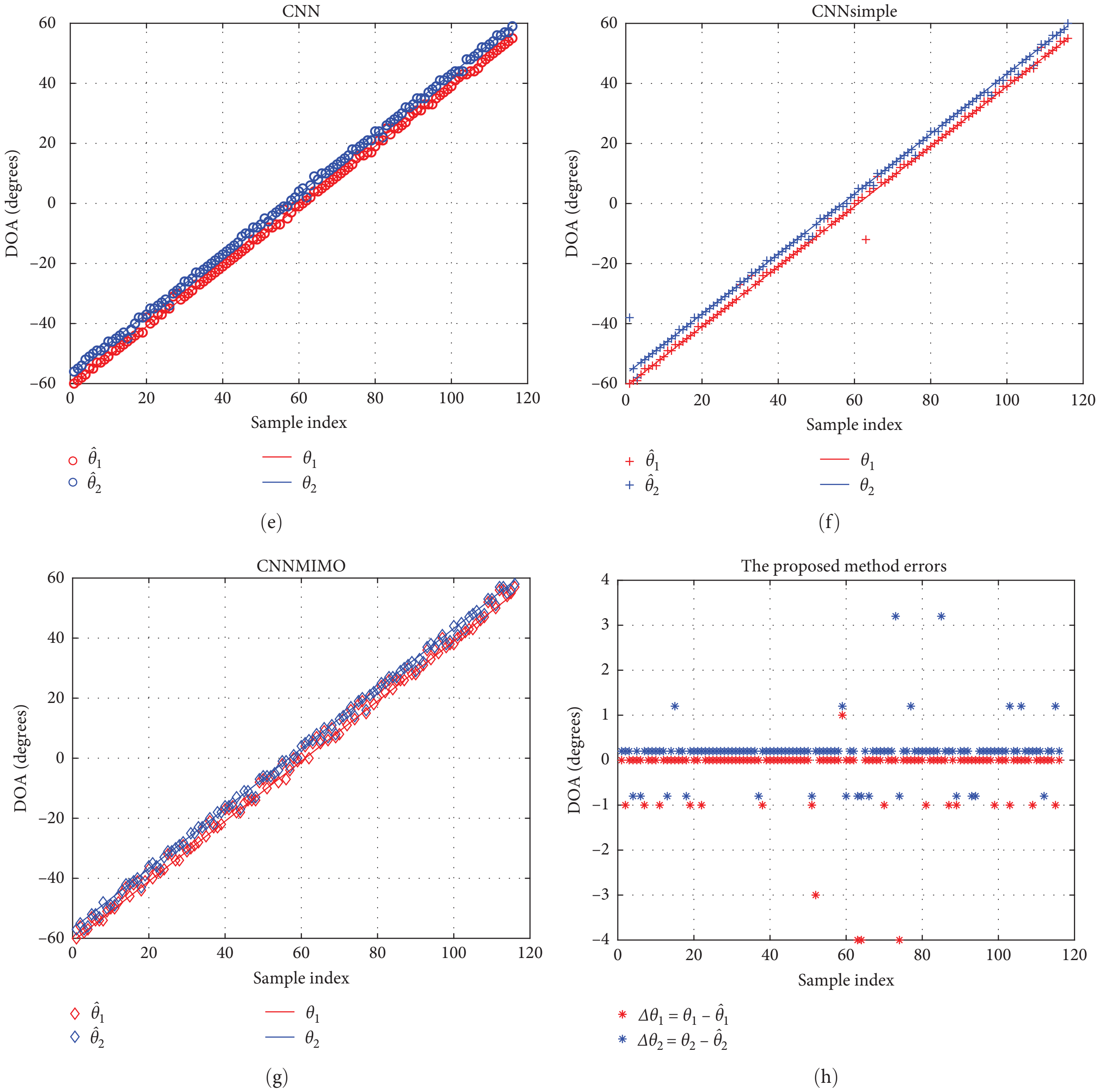
(a)

(b)

(c)

(d)

Figure 5: Continued.



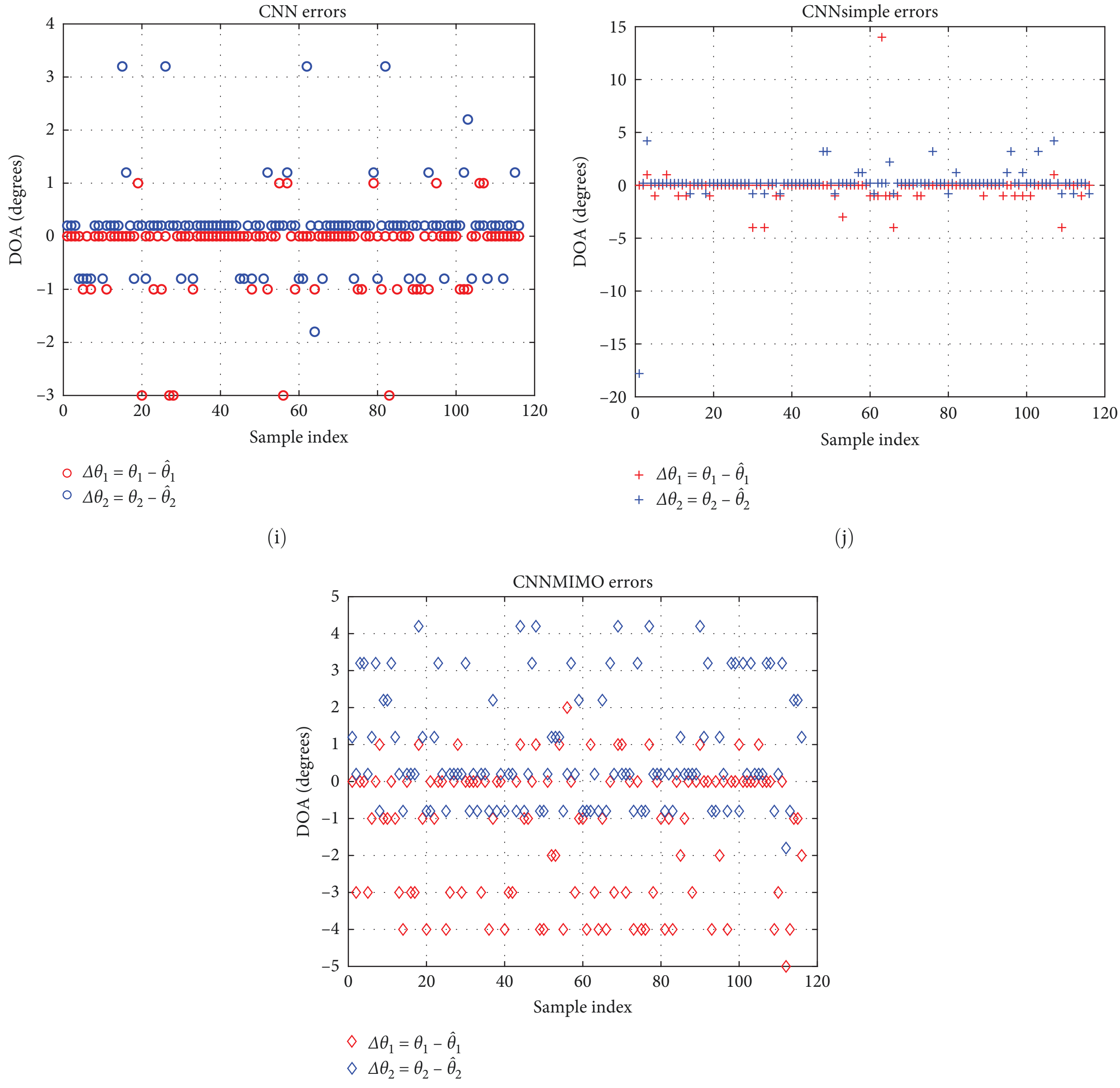
(e)

(f)

(g)

(h)

FIGURE 5: Continued.



(i)

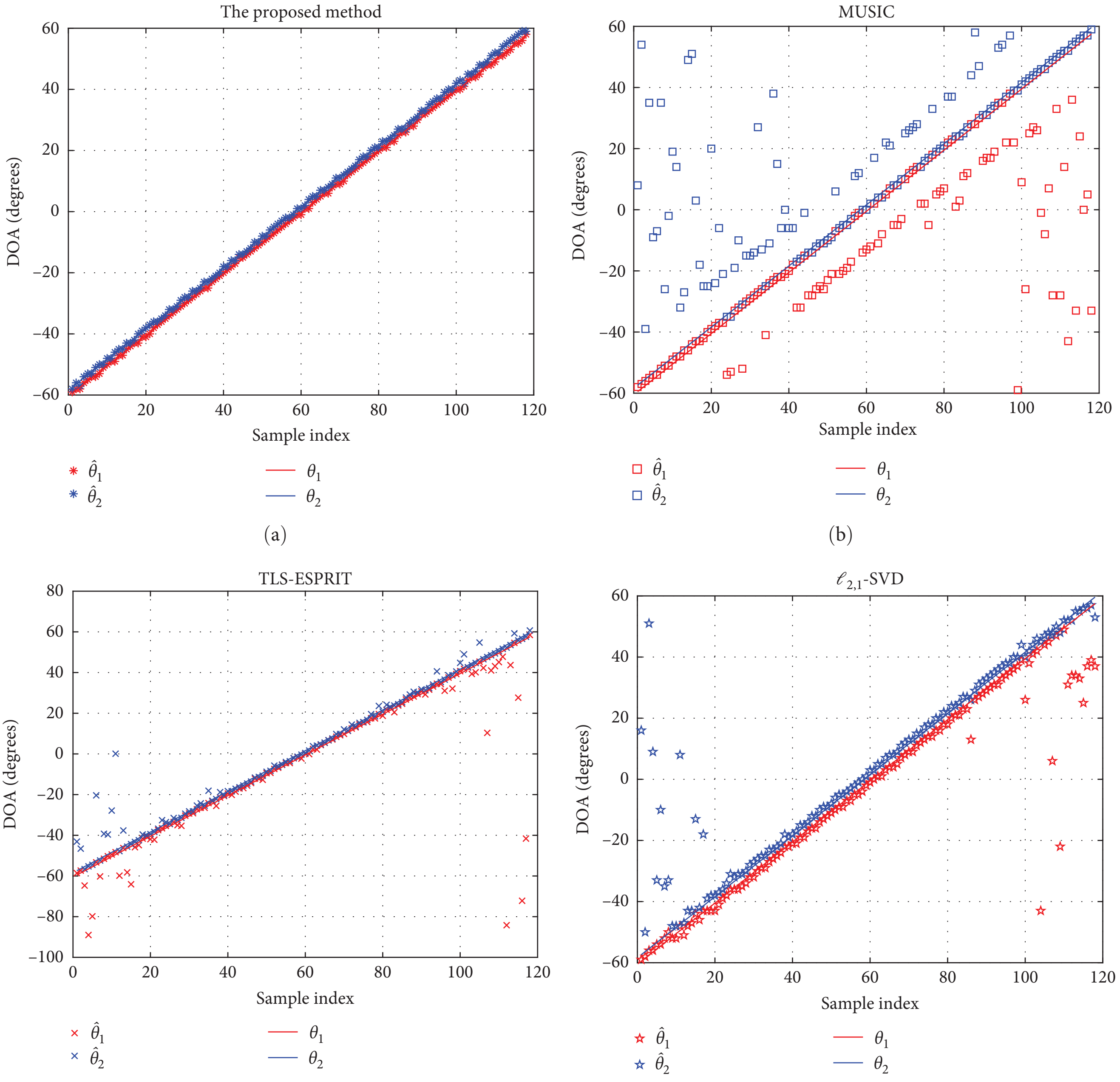
(j)

(k)

Figure 5: DOA estimates at the off-grid angles $\theta_1, \theta_2 \in [-60^\circ, 60^\circ]$ for −10 dB and 500 snapshots. The DOA estimates of (a) the proposed method, (b) MUSIC, (c) TLS-ESPRIT, (d) $\ell_{2,1}$-SVD, (e) CNN, (f) CNNsimple, and (g) CNNMIMO. The DOA estimation errors of (h) the proposed method, (i) CNN, (j) CNNsimple, and (k) CNNMIMO.

included in the training set, to verify its effect in the case of off-grid angles. Figure 5 displays the DOA estimates of different algorithms and their deviations from the real angle estimates when two different sources are separated by a certain angle. Among them, Figure 5(a)–5(g) show the DOA estimation result of the proposed model, MUSIC, TLS-ESPRIT, $\ell_{2,1}$-SVD, CNN, CNNsimple, and CNNMIMO, respectively. Figure 5(h)–5(k) show the DOA estimation error of the proposed algorithm, CNN, CNNsimple, and CNNMIMO, respectively.

In Figure 5, each angle of θ_1 and θ_2 is connected by a straight line, which makes it more intuitive to see how much the angle pairs are offset when the estimate is wrong (in future such experiments, all angles of θ_1 and θ_2 will also be connected separately as lines). It can be seen that the methods using neural networks are more effective (proposed method, CNN, CNNsimple, and CNNMIMO). The MUSIC algorithm has limitations in angle prediction. It can accurately predict only one angle at the edge, but struggles to estimate two angles in the remaining range, resulting in significant errors. TLS-ESPRIT also has difficulty separating angles correctly, but it can predict angles well in most ranges. In the edge angle range, it has a higher probability of separating only one angle, but its performance is still better



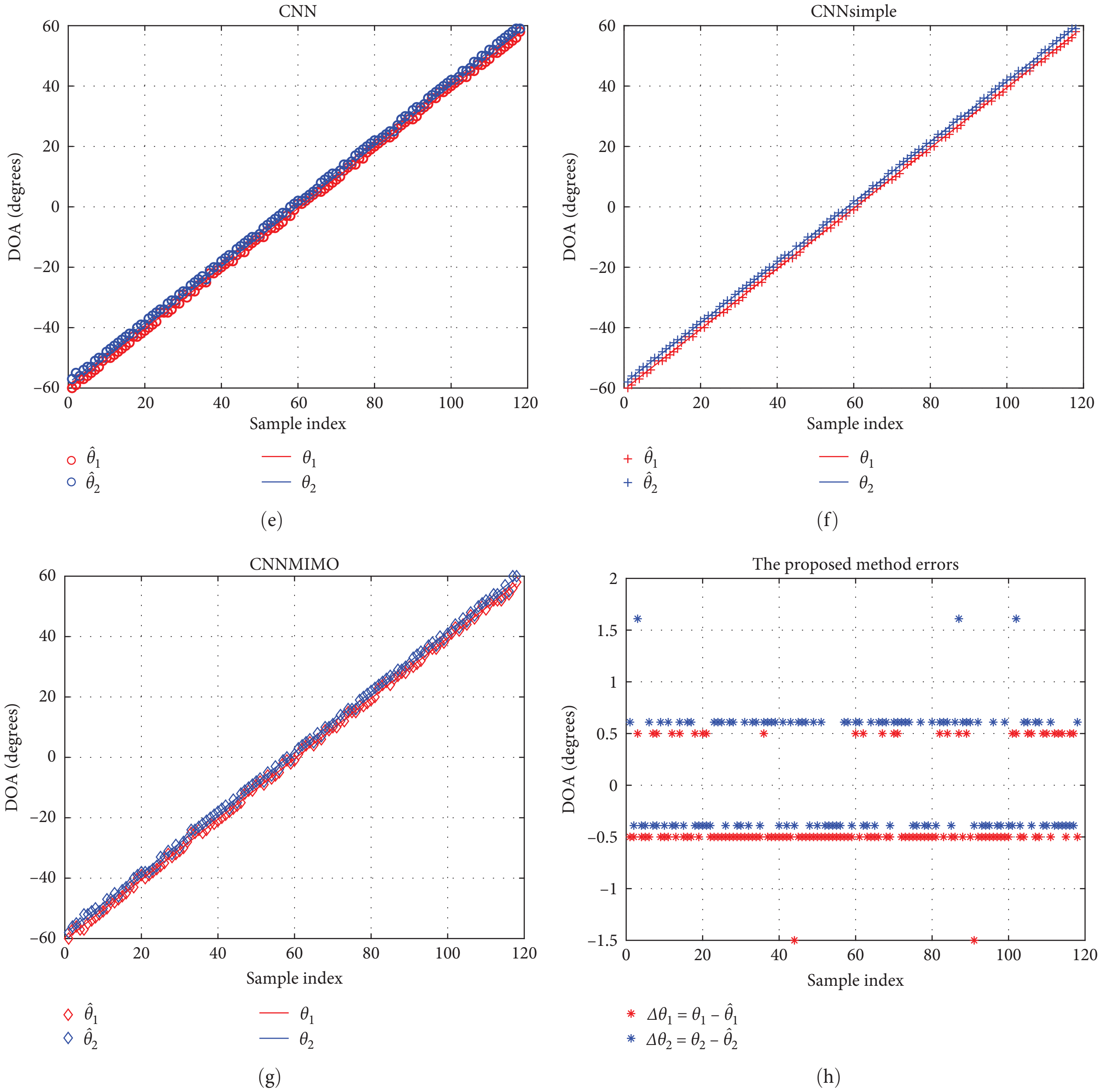
(a)

(b)

(c)

(d)

Figure 6: Continued.



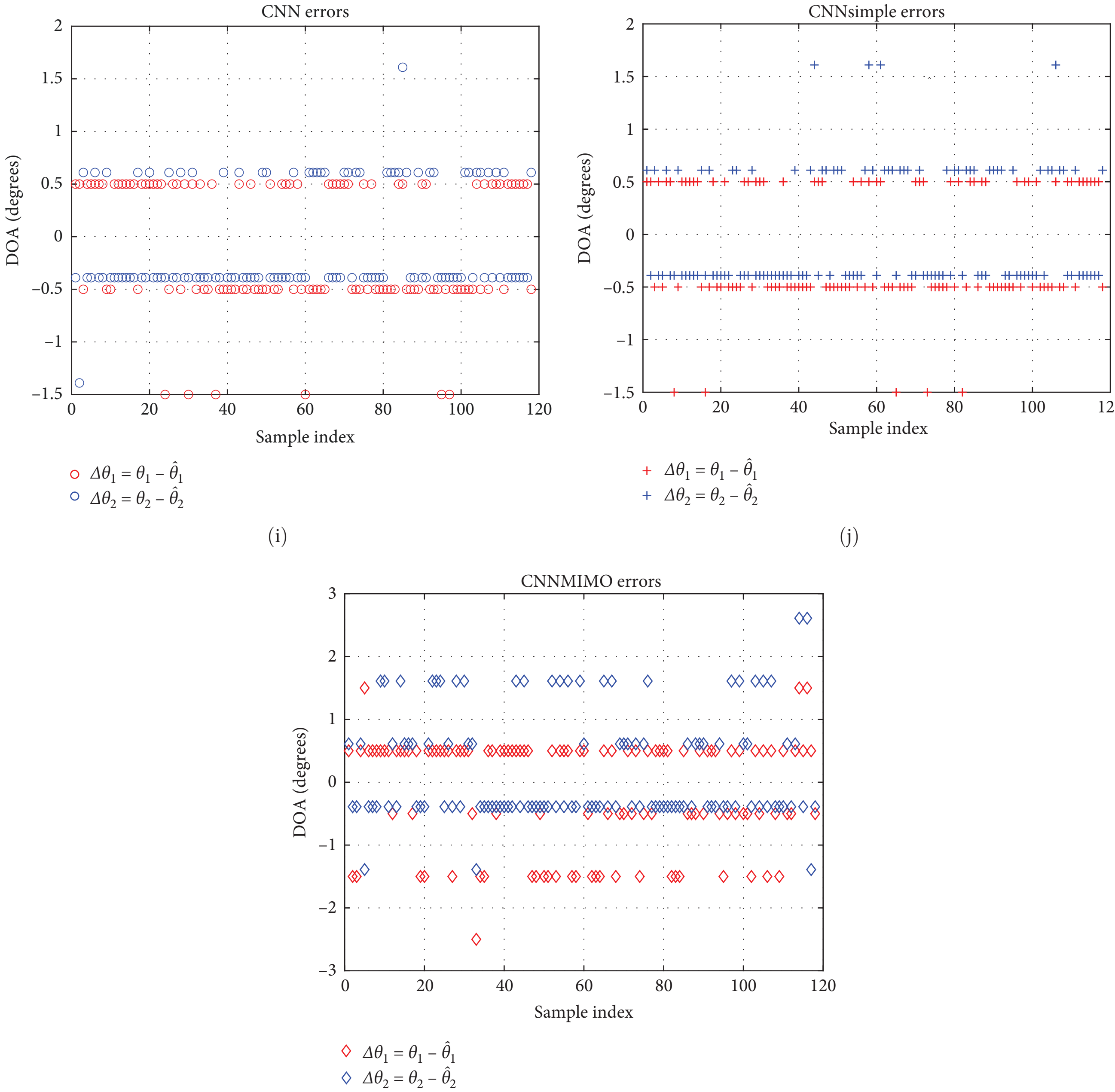
(e)

(f)

(g)

(h)

Figure 6: Continued.



(i)

(j)

(k)

Figure 6: DOA estimates at the off-grid angles $\theta_1, \theta_2 \in [-60^\circ, 60^\circ]$ for -5 dB and 400 snapshots. The DOA estimates of (a) the proposed method, (b) MUSIC, (c) TLS-ESPRIT, (d) $\ell_{2,1}$-SVD, (e) CNN, (f) CNNsimple, and (g) CNNMIMO. The DOA estimation errors of (h) the proposed method, (i) CNN, (j) CNNsimple, and (k) CNNMIMO.

than that of MUSIC. Similar to TLS-ESPRIT, $\ell_{2,1}$-SVD also has the problem that there is a probability that only one edge angle can be measured. The difference between the proposed model and the CNN model is small. The former is [−4°, 3.2°], and the latter is [−3°, 3.2°]. For the RMSE index of performance evaluation, the RMSE of the proposed method is 0.7054 and 23.3266 for MUSIC, 8.9719 for TLS-ESPRIT, 3.2772 for $\ell_{2,1}$-SVD, 0.8058 for CNN, 1.7612 for CNNsimple, and 1.9892 for CNNMIMO. Therefore, under the conditions of a certain source distance, low SNR and a reasonable number of snapshots, the proposed model has a better effect than others. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 290$.

Next, consider the second set of experiments. In this set of experiments, the source spacing is close and the DOA is off-grid. The SNR was -5 dB, the number of snapshots was 400, and the angular spacing used was $\Delta\theta = 2.11^\circ$. The first signal, θ_1, starts at -59.5° in steps of 1° and ends at 57.5°. Similarly, for each θ_1, $\theta_2 = \theta_1 + \Delta\theta$. Figure 6 shows the DOA estimates of different algorithms and their deviations from the real angles when two different sources are very close to each other and the angles are not on the grid. As in the

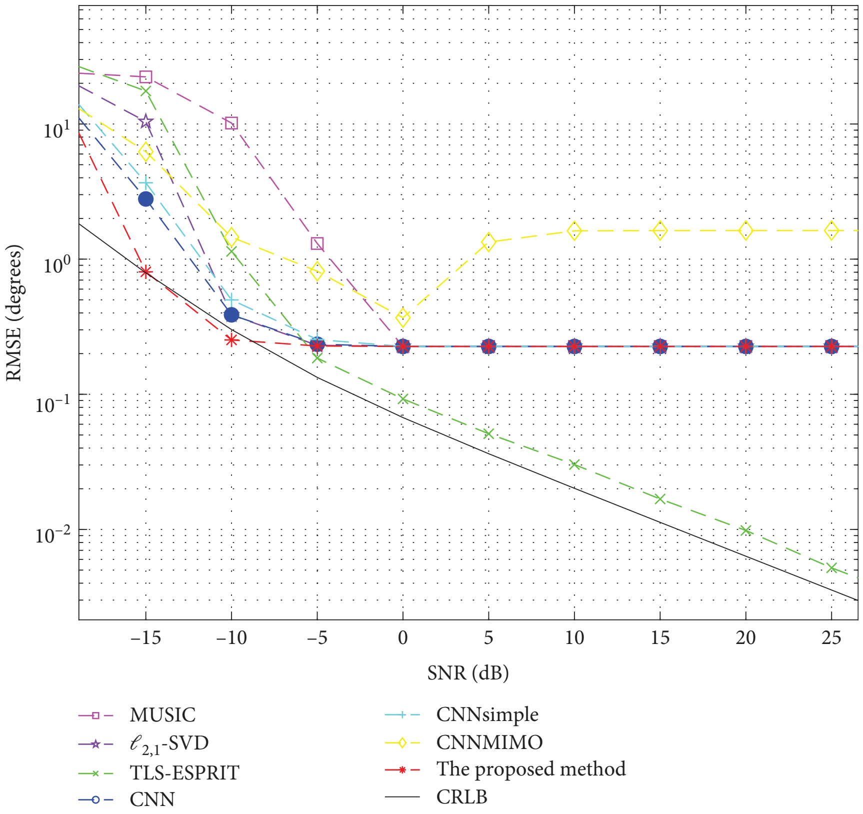


FIGURE 7: The RMSE of DOA estimation for two off-grid angle sources under different SNRs when the number of snapshots used by different methods is 1,000.

previous experiment, Figure 6(a)–6(g) show the DOA estimation result of the proposed model, MUSIC, TLS-ESPRIT, $\ell_{2,1}$-SVD, CNN, CNNsimple, and CNNMIMO, respectively. Figure 6(h)–6(k) show the DOA estimation error of the proposed algorithm, CNN, CNNsimple, and CNNMIMO, respectively.

For the 1° resolution method, the errors should be $\{-0.5^\circ, 0.5^\circ\}$ and $\{-0.39^\circ, 0.61^\circ\}$ for grids with similar angles. It can be seen that when the number of snapshots is small and the DOA of the source is similar, only the DL method can estimate the DOA effectively, whereas the other methods cannot estimate the DOA well. The reasons for the poor estimation of the other methods are as follows: the other methods cannot separate the two sources that are relatively close and treat the two sources as one; or the angle is separated, but the angle is still similar, so the predicted angle and the actual angle have a small deviation. MUSIC cannot separate angles. Although ESPRIT and $\ell_{2,1}$-SVD can effectively separate two sources, there is a probability that one of them will fail to be estimated, resulting in some error, so their RMSE is slightly larger than that of the DL algorithms. CNN, CNNsimple, and the proposed model can accurately estimate the angle within the error of $[-1.5^\circ, 1.61^\circ]$, indicating the effectiveness of the DL method in DOA estimation when the number of snapshots is small and the source is similar. The RMSE of the different methods is as follows: 0.5497 for the proposed method, 26.3381 for MUSIC, 15.0796 for TLS-ESPRIT, 13.9925 for $\ell_{2,1}$-SVD, 0.5617 for CNN, 0.5700 for CNNsimple, and 0.8956 for CNNMIMO. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 140$. The proposed method still has good performance with smaller SNR, small angular spacing, and less number of snapshots.

5.3. DOA Estimation at Different SNRs. This section examines the performance of DOA estimation at various SNRs. In the experiment, the number of snapshots used is 1,000, while the SNRs used range from -20 to 30 dB with a step size of 5. The two angles used are $\theta_1 = 10.11^\circ$ and $\theta_2 = 13.3^\circ$. For these two signal sources, 500 times Monte Carlo experiments were performed with different algorithms at different SNRs to explore the DOA estimation performance of each algorithm, and the results are shown in Figure 7. As shown in Figure 7, the proposed method, CNN, and CNNMIMO have good results under the condition of low SNRs (roughly from -15 to -5 dB). In the case of medium and high SNRs (roughly from -5 to 30 dB), the on-grid methods begin to be inferior to the off-grid methods, and the gap in estimation accuracy becomes larger and larger. When the on-grid method reaches the minimum error caused by the grid, the RMSE does not decrease with an increase in SNR (in the case of $\theta_1 = 10.11^\circ$ and $\theta_2 = 13.3^\circ$, the minimum error value of this RMSE is 0.2259). This error is insurmountable and can only be reduced by reducing the grid spacing to obtain a finer grid. Because the off-grid method overcomes the limitations of the grid, the RMSE can continuously decrease with the increase in SNR, which has better performance under the

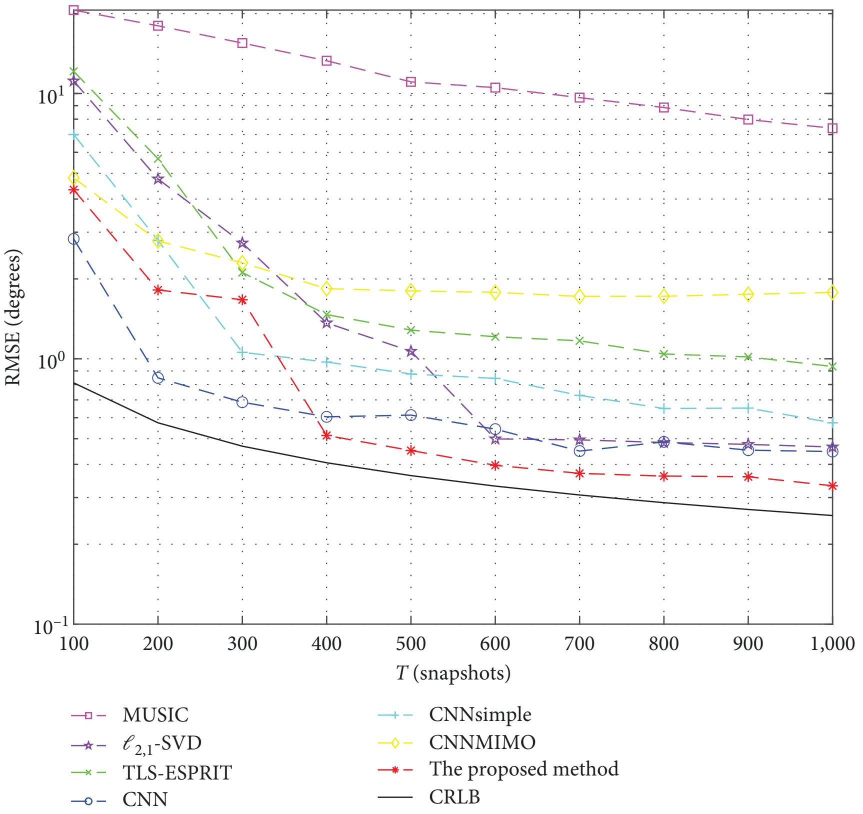


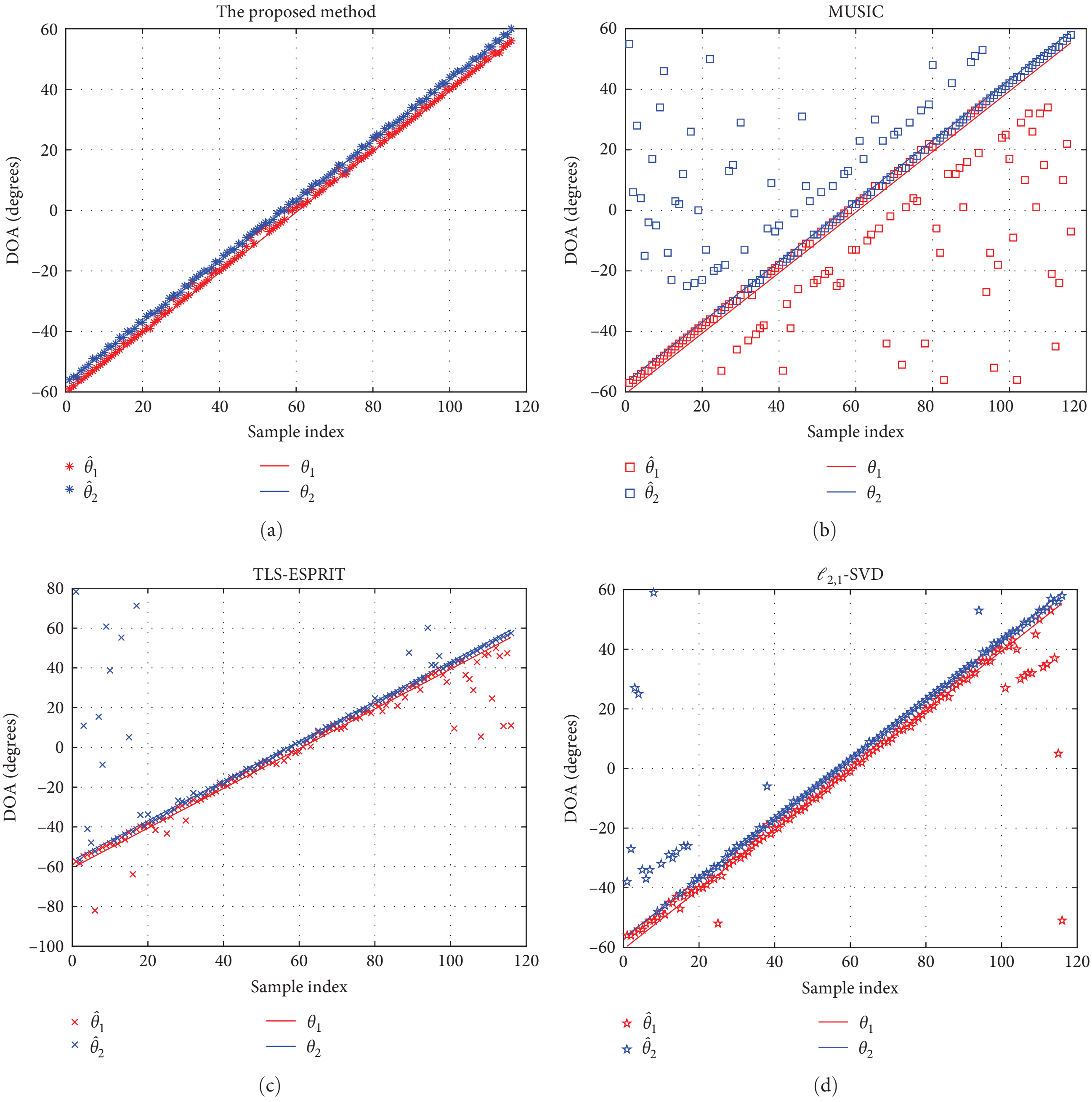
FIGURE 8: The RMSE of DOA estimation for two off-grid angle sources with different numbers of snapshots and the SNR of -10 dB by different methods.

condition of high SNR. The effectiveness of CNNMIMO training methods may be limited to low SNR and reduced at high SNR. Furthermore, modifying the amount of training data to be of the same order of magnitude as other DL algorithms may have decreased their effectiveness at low SNR. The RMSE of the proposed method is lower than that of CRLB at -10 dB, because the DL method is not an unbiased estimation of DOA like other algorithms, it is obtained by training, so it is a biased estimation, and its performance can be better than the unbiased estimation in some cases. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = \{1270.720, 400, 230, 130, 70, 40, 20, 20, 20, 20\}$ as the SNR changes.

5.4. DOA Estimation for Different Number of Snapshots. In experiments with different numbers of snapshots, the SNR of the two sources is -10 dB, the number of snapshots is chosen from 100 to $1,000$, and the step size used is 100. The θ_1 used is -14.77° and the θ_2 used is -11.17°. For the two sources above, Monte Carlo experiments were run 500 times using different methods with different numbers of snapshots. The RMSE results obtained by different methods using different numbers of snapshots are shown in Figure 8. As shown in Figure 8, the proposed method has the smallest RMSE when the number of snapshots is between 400 and 1,000, whereas the CNN has the best performance when the number of snapshots is between 100 and 300. The performance of CNNsimple is between CNN and TLS-ESPRIT and outperforms the proposed algorithm at a snapshot count of 300. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = \{130, 180, 220, 260, 290, 310, 340, 360, 380, 400\}$ as the number of snapshots changes. The proposed algorithm is more advantageous when dealing with a large number of snapshots.

5.5. DOA Estimation Under Signal Mismatch. In previous experiments, the DOA estimation performance of each algorithm studied was based on the same power of two sources. In reality, however, they are often not the same. Repeat the two DOA estimation experiments in Section 5.2 but change some of the conditions. In the following two experiments, change the power of two different sources to $\sigma_1^2 = 0.7$, $\sigma_2^2 = 1.25$ and $\sigma_1^2 = 1.25$, $\sigma_2^2 = 0.7$, respectively, in the following two experiments. σ_1^2 represents the power of θ_1, i.e., the power of the smaller angle, whereas σ_2^2 represents the power of θ_2, i.e., the power of the larger angle. In both cases, the SNR is 1.549 dB higher in the case of a signal match than in the case of no match. The experiment used two different power forms of signal sources in two conditions: a larger angle interval with low SNR and more snapshots and a smaller angle interval with higher SNR and fewer snapshots.

The first experiment differs from the first experiment in Section 5.2. The SNR is -10 dB, the number of snapshots is 400, θ_1 starts at -59.63° with a step size of 1° to 55.37°, and the interval between θ_2 and θ_1 is 3.7°. The first type of mismatch mentioned above is selected. The DOA estimation and the DOA estimation error of the different methods in



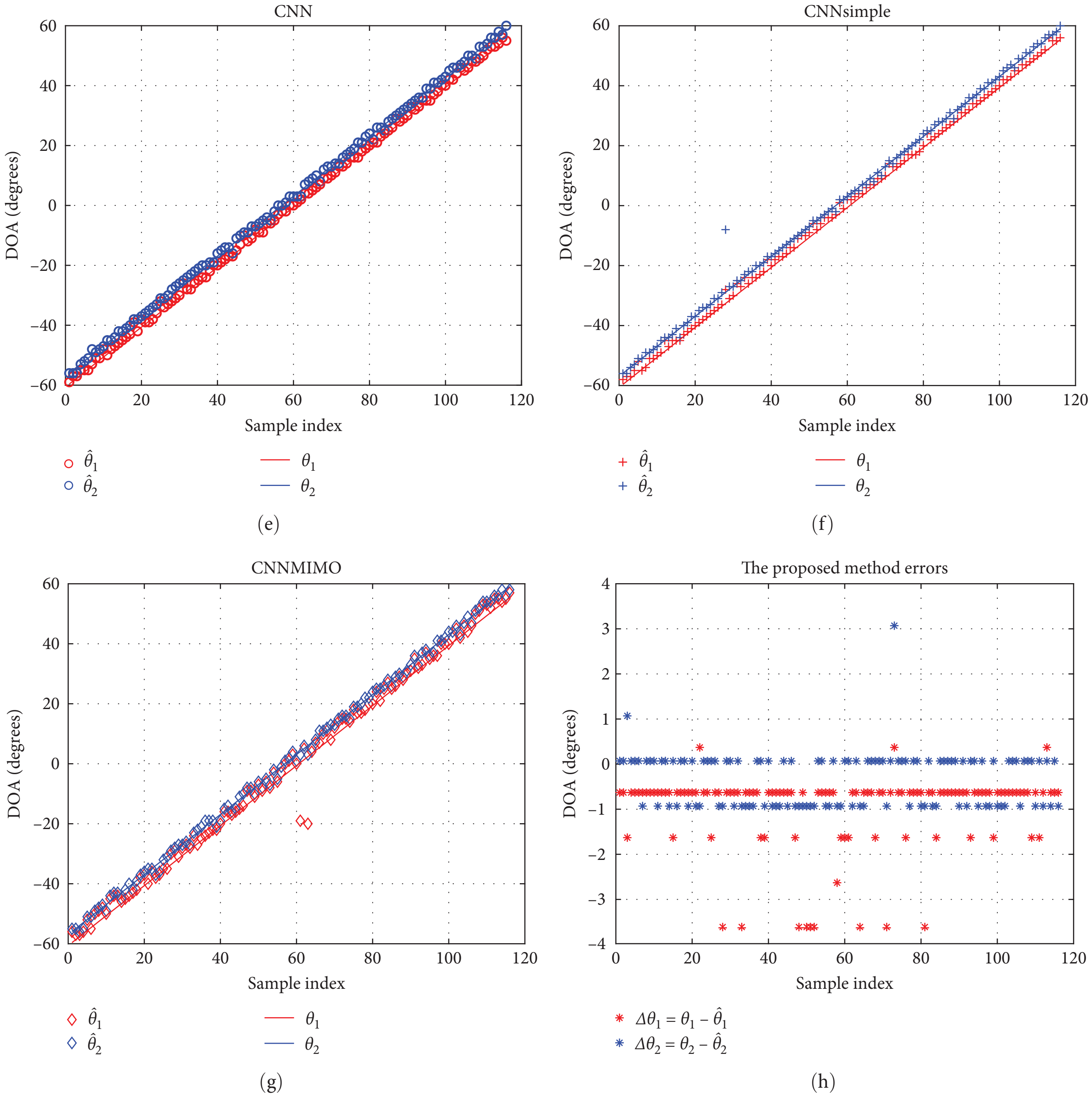
(a)

(b)

(c)

(d)

Figure 9: Continued.



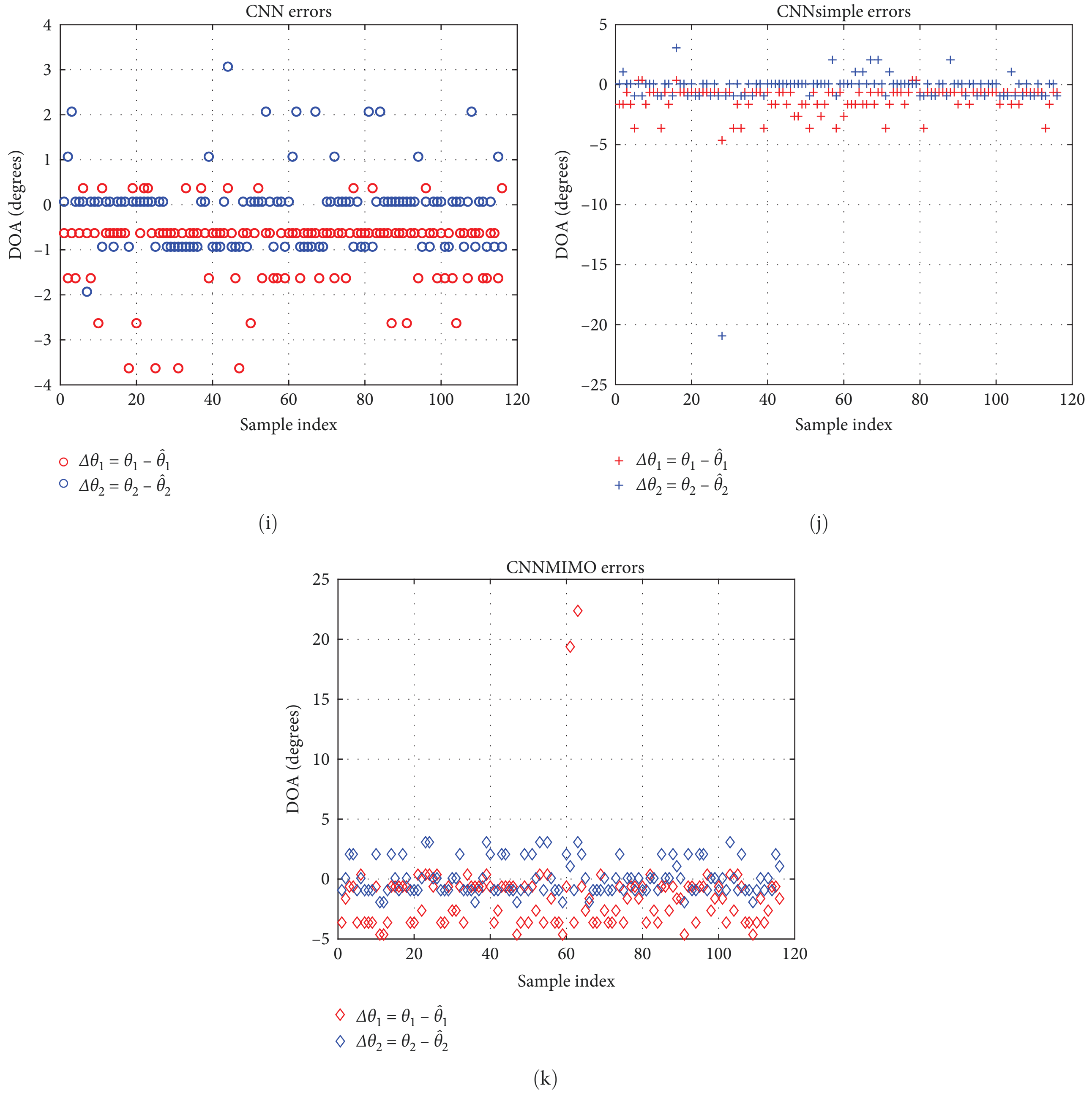
(e)

(f)

(g)

(h)

Figure 9: Continued.



(i)

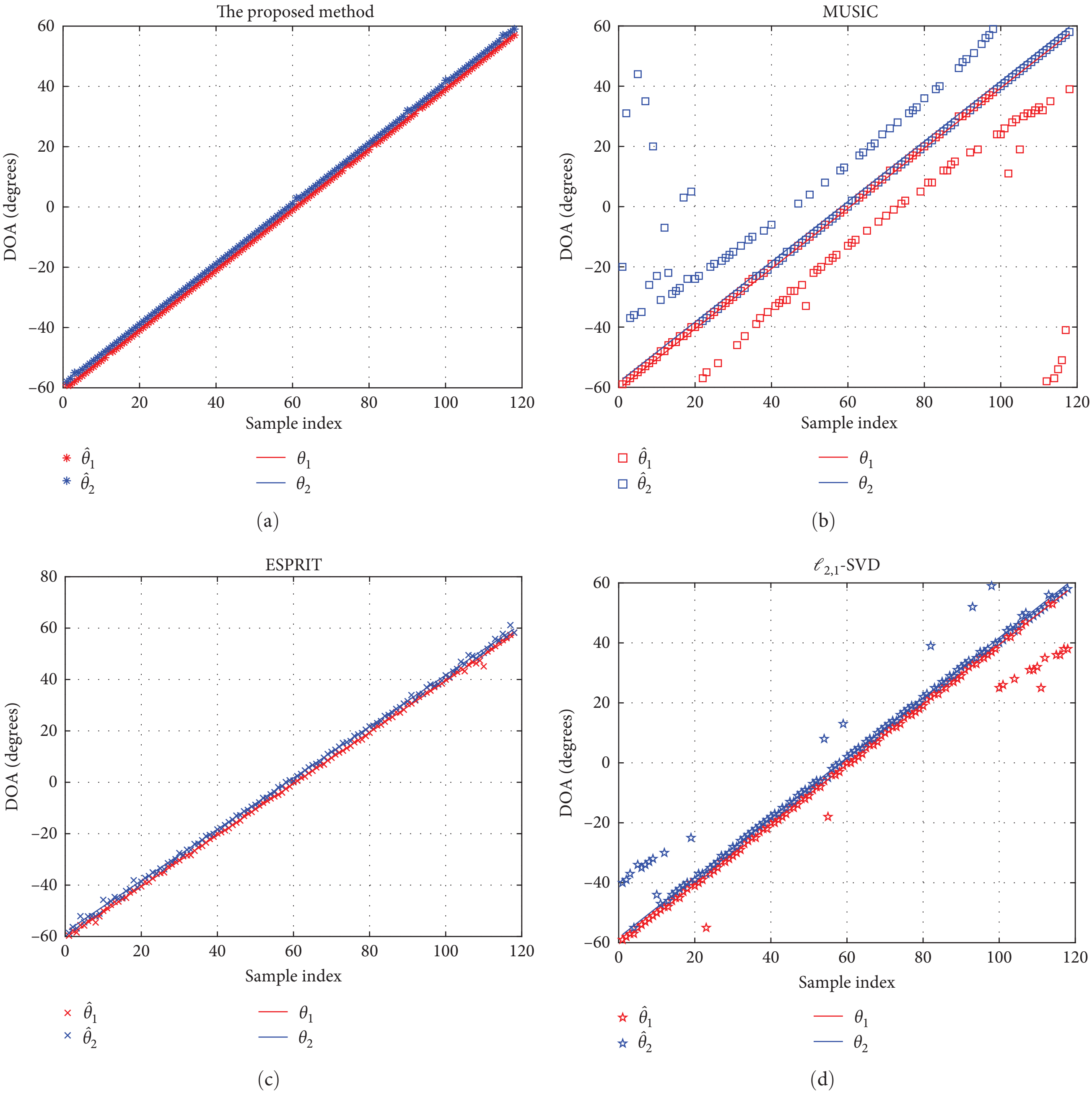
(j)

(k)

Figure 9: DOA estimates at the off-grid angles $\theta_1, \theta_2 \in [-60^\circ, 60^\circ]$ for −10 dB and 400 snapshots. The DOA estimates of (a) the proposed method, (b) MUSIC, (c) TLS-ESPRIT, (d) $\ell_{2,1}$-SVD, (e) CNN, (f) CNNsimple, and (g) CNNMIMO. The DOA estimation errors of (h) the proposed method, (i) CNN, (j) CNNsimple, and (k) CNNMIMO. The proposed method has the best performance with an RMSE of 1.0544.

this experiment are shown in Figure 9. In the case of a signal power mismatch, MUSIC can still only predict one angle, and the predicted angle is not related to the signal power. Therefore, the RMSE is very high. TLS-ESPRIT and $\ell_{2,1}$-SVD also suffer from the edge angle prediction bias mentioned in Section 5.2. CNNsimple and CNNMIMO have one or two angles with large estimation errors. In this case, CNN and the proposed algorithm have excellent performance, and they can all predict the DOA well in the case of two sources. They both have an error of $[-3.63^\circ, 3.07^\circ]$. The RMSE of the different methods is as follows: 1.0544 for the proposed method, 27.6415 for MUSIC, 18.8163 for TLS-ESPRIT, 13.7638 for $\ell_{2,1}$-SVD, 1.0710 for CNN, 1.8402 for CNNsimple, and 2.7700 for CNNMIMO. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 260$.

The second experiment differs from the second experiment in Section 5.2. The SNR is 0 dB, the number of snapshots is 200, θ_1 starts at -59.6° with steps of 1° to 57.4°, and angular intervals of 2.3°. The second form of mismatch mentioned above is selected. The DOA estimation results and DOA estimation errors of this experiment are shown in Figure 10. In this case, MUSIC still has the problems



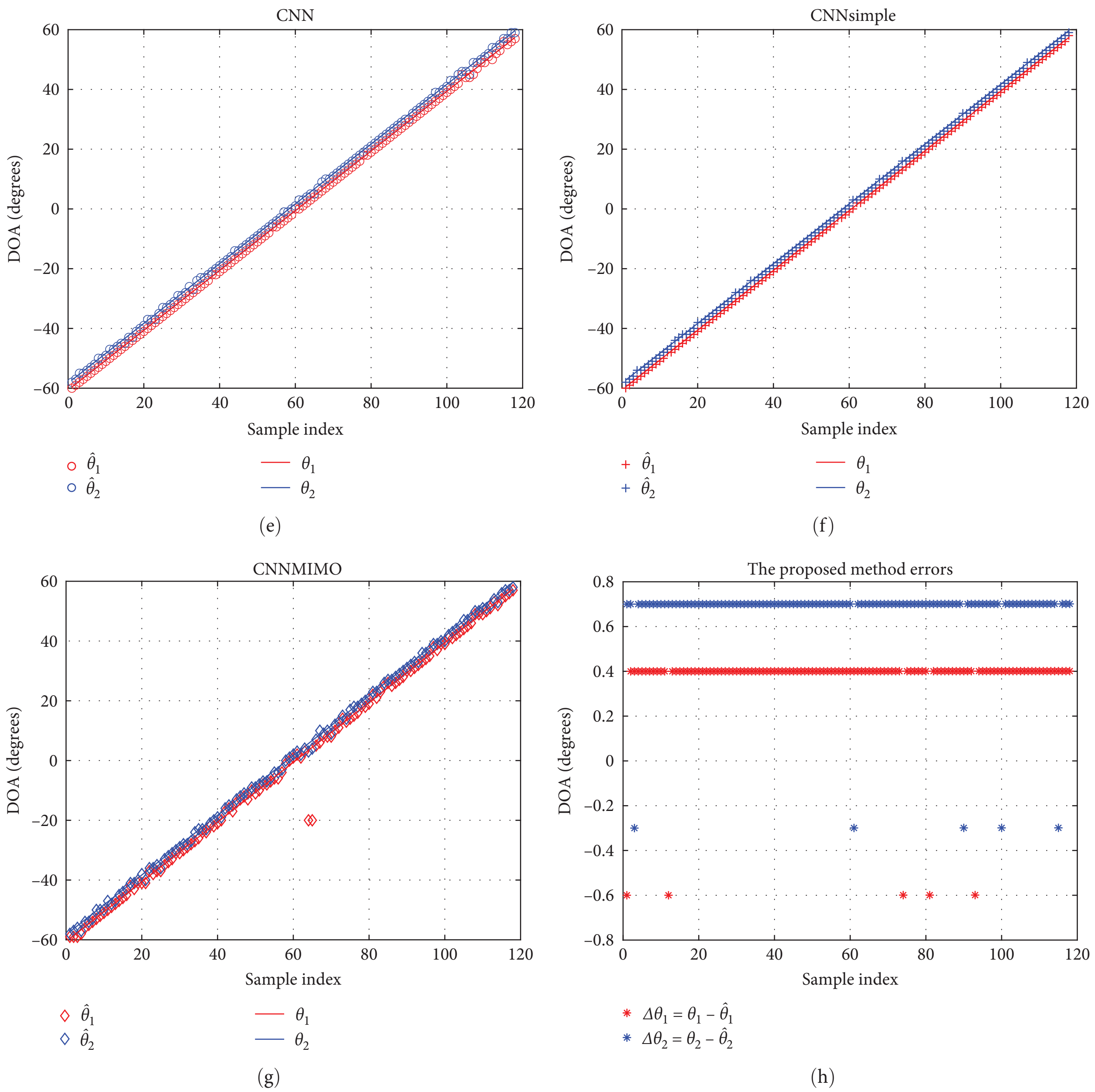
(a)

(b)

(c)

(d)

Figure 10: Continued.



(e)

(f)

(g)

(h)

Figure 10: Continued.

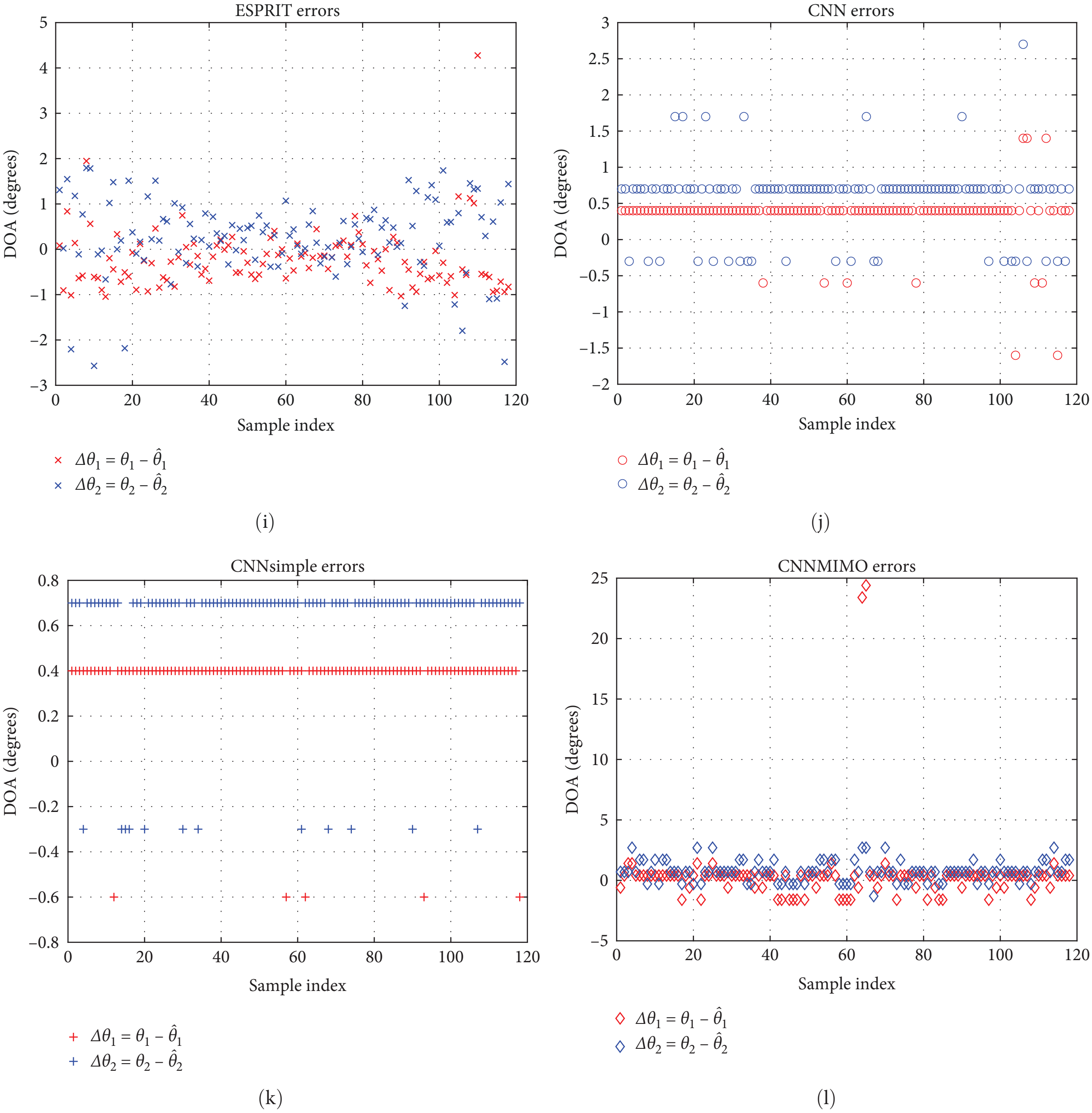


FIGURE 10: DOA estimates at the off-grid angles $\theta_1, \theta_2 \in [-60^\circ, 60^\circ]$ for 0 dB and 200 snapshots. The DOA estimates of (a) the proposed method, (b) MUSIC, (c) TLS-ESPRIT, (d) $\ell_{2,1}$-SVD, (e) CNN, (f) CNNsimple, and (g) CNNMIMO. The DOA estimation errors of (h) the proposed method, (i) ESPRIT, (j) CNN, (k) CNNsimple, and (l) CNNMIMO.

mentioned in the previous experiment. The $\ell_{2,1}$-SVD algorithm can only estimate one angle in the edge angle. CNNMIMO still suffers from a few estimation errors. These problems with the above algorithms cause their RMSE to be too high, whereas TLS-ESPRIT, CNN, CNNsimple, and the algorithm proposed in this paper have excellent performance. The DOA estimation error ranges of TLS-ESPRIT and CNN are $[-2.57^\circ, 4.275^\circ]$ and $[-1.6^\circ, 2.7^\circ]$, respectively, whereas the DOA estimation error ranges of this paper's method and CNNsimple are only $[-0.6^\circ, 0.7^\circ]$. The RMSE of the different methods is as follows: 0.5664 for the proposed method, 22.2985 for MUSIC, 0.7940 for TLS-ESPRIT, 6.0440 for $\ell_{2,1}$-SVD, 0.6526 for CNN, 0.5558 for CNNsimple, and 2.4063 for CNNMIMO. The proposed method has a relatively small estimation error and an RMSE second only to CNNsimple. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 60$.

The above two experiments show that the proposed method can perform better DOA estimation than other methods under the condition of signal power mismatch.

5.6. DOA Estimation at Different Separation Angles. Previous experiments have all considered the case of constant source separation angle. This section discusses the DOA estimation

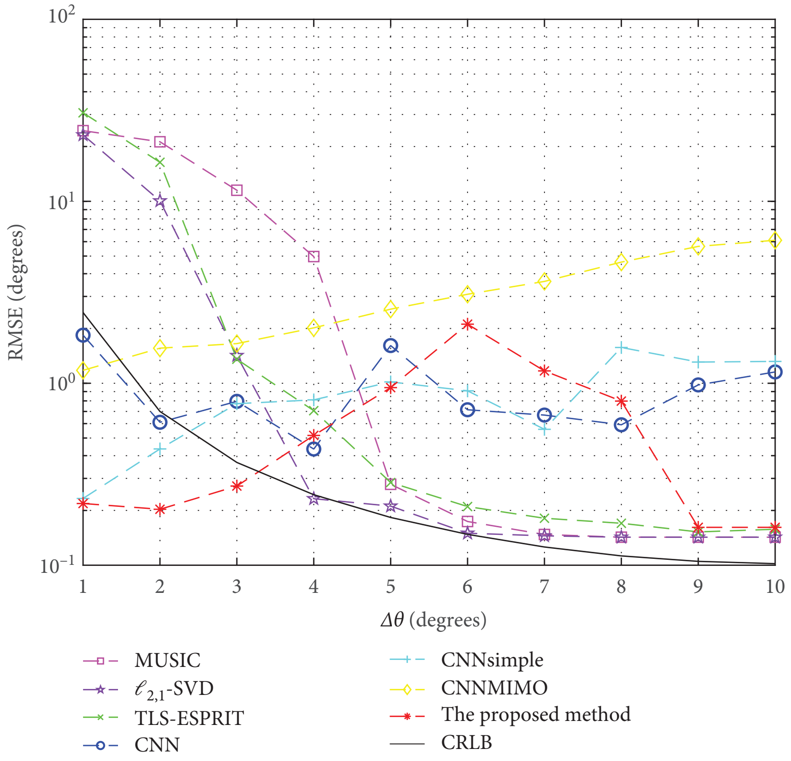


Figure 11: The RMSE of various algorithms under 500 times monte carlo experiments, when the SNR is −10 dB, the number of snapshots is 800, and the separation angle is increased from 1° to 10° with a step size of 1°.

performance of different algorithms under different separation angles. The SNR chosen for both signals is −10 dB and the number of snapshots is 800. The θ_1 is chosen to be -12.14° and the angular separation $\Delta\theta$ is increased from 1° to 10° with step size of 1°. The RMSE of different algorithms under different $\Delta\theta$ was obtained by Monte Carlo experiments for 500 times, as shown in Figure 11.

As shown in Figure 11, the proposed algorithm has better performance at different separation angles in most cases. However, the RMSE is slightly higher for separation angles from 5° to 7°. In the case of small separation angles (separation angles from 1° to 3°), the best performing methods are those of the DL algorithms (CNN, CNNsimple, CNNMIMO, and the proposed methods), and their RMSE is lower than that of CRLB because they are biased estimates. CNN has a better performance than the proposed method in the interval of separation angle from 6° to 8° and the separation angle is 4°, and other different separation angles are inferior to the proposed method. CNNsimple outperforms the proposed algorithm in the 6° to 7° interval. The increased RMSE of CNNMIMO with separation angle may be due to fewer training samples. Compared with other algorithms (MUSIC, TLS-ESPRIT, and $\ell_{2,1}$-SVD), the proposed method has a better performance in the separation angle from 1° to 3°, and the RMSE of these methods is larger. These methods all have small RMSE when the separation angle is large. Although the performance of the proposed method is inferior to those of these algorithms when the separation angle is large, the proposed algorithm still has a relatively suitable RMSE at these separation angles, and will not have a very large RMSE when the separation angle is small. The proposed method is more suitable in the case of similar sources, and can also give a better estimate of sources that are not close. When the separation angle is larger than 8°, the proposed algorithm can predict the DOA as well as other algorithms (MUSIC, TLS-ESPRIT, and $\ell_{2,1}$-SVD). However, the RMSE of the CNN algorithm does not decrease at larger separation angles, i.e., it has a larger RMSE at different separation angles. MUSIC, TLS-ESPRIT, and $\ell_{2,1}$-SVD estimate DOA well for separation angles greater than 5° but are particularly poor for closer sources. It is worth noting that another biased estimate of the $\ell_{2,1}$-SVD, the RMSE is lower than that of the CRLB when the separation angle is 4°. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 360$.

5.7. Signal-Dependent DOA Estimation. Previous experiments considered the case of noncorrelation of source signals. This section discusses the case of signal correlation and shows that the proposed method still has good performance in the case of signal correlation. In this case, the chosen source angle is $\theta_1 = 11.13^\circ$ and $\theta_2 = 14.45^\circ$. Note that in this case, the diagonal source covariance matrix $\mathbf{R}_s$ becomes $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where ρ is the correlation coefficient of the two signal sources. A low SNR is still used when $\mathbf{R}_s$ changes. The selected SNR is −10 dB and the number of snapshots is 500. Under the condition that the step size of the correlation coefficient is 0.1, Monte Carlo experiments of different

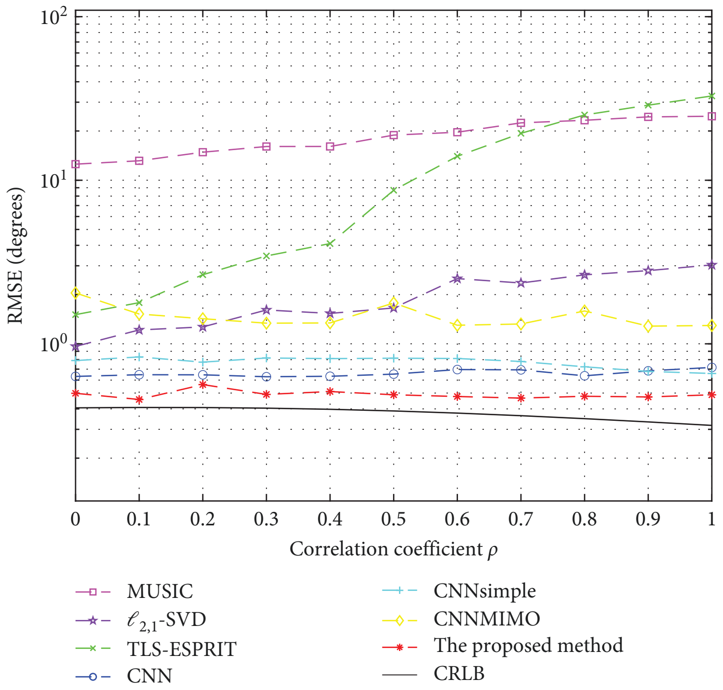


Figure 12: The RMSE of different algorithms varies with the correlation coefficient when the SNR is -10 dB and the number of snapshots is 500. The step size of the correlation coefficient is 0.1.

Table 2: Processing time for algorithm.

Algorithm	Processing time
MUSIC	1.364 ms
TLS-ESPRIT	4.731 ms
$\ell_{2,1}$-SVD	39.204 s
CNN	2.667 ms
CNNsimple	2.002 ms
CNNMIMO	1.805 ms
The proposed method	0.772 ms

Table 3: Source number uncertainty DOA estimation accuracy and Hausdorff distance.

Algorithm	K	Mean d_H	Max d_H	Accuracy (%)
CNN	1	0.40	0.40	100
	2	0.34	4.8	99.84
	3	4.1	9.8	86.32
The proposed method	1	0.40	0.40	100
	2	0.37	5.3	99.80
	3	3.2	4.4	92.85

algorithms are performed for 500 times with the correlation coefficient ranging from 0 to 1, and the RMSE obtained is shown in Figure 12. Figure 12 demonstrates the robustness of DL algorithms to variations in ρ. The performance of the DL algorithms, from worst to best, is CNNMIMO, CNNsimple, CNN, and the algorithm proposed in this paper. The threshold of the $\ell_{2,1}$-SVD in this experiment is $\eta = 290$.

When ρ equals 0, the efficiency of processing different datasets is compared by measuring the processing time of each algorithm. Table 2 presents the results, which show that the method has the shortest processing time.

5.8. DOA Estimation with Unknown Number of Sources. When K is not equal to 2, the same training method was used to obtain a model that can adapt to different numbers of sources. The model now includes source number discrimination to handle cases where the source number is unknown. The number of neurons in layer FC3 in Figure 3 was changed to $K-1$, i.e., 15. This enables the network to transition from differentiating between angles to differentiating between the quantity of sources by identifying the index with the highest value as the number of sources. In the tests, $K_{\max} = 3$ was chosen as the maximum number of sources. Because there may be cases where the number of sources is not judged correctly, resulting in a nonadaptive RMSE, the Hausdorff distance is used to evaluate the difference in results with the following formula:

$$d_H(\mathscr{A}, \mathscr{B}) = \max\{d(\mathscr{A}, \mathscr{B}), d(\mathscr{B}, \mathscr{A})\}, \tag{20}$$

where

$$d(\mathscr{A}, \mathscr{B}) = \sup\{d(\alpha, \mathscr{B}) | \alpha \in \mathscr{A}\}, \tag{21}$$

$$d(\alpha, \mathscr{B}) = \inf\{d(\alpha, \beta) | \beta \in \mathscr{B}\}, \tag{22}$$

$$d(\alpha, \beta) = |\alpha - \beta|. \tag{23}$$

The signal is -5.8° at $K = 1$, 3.3° is added to the signal at $K = 2$, and 8.4° is added to the signal at $K = 3$. Table 3 presents the accuracy and Hausdorff distance of both the CNN and the proposed algorithm for the selected SNR of 0 dB and 1,000 snapshots (10,000 experiments). It can be seen that our algorithm has performance similar to the CNN method.

6. Conclusion

In this paper, transformer is applied to the field of DOA estimation for the first time, and a transformer-based DOA estimation model that can be adapted to the low SNR situation is proposed. The problem is modelled as a multilabel classification model of on-grid angles. The improved MHA is used to extract features from the processed multichannel data, and the original Star-Transformer is improved. A robust DOA estimation model can be obtained through a series of subsequent processing steps. In future studies, we aim to improve the robustness of the model in real-world scenarios by adjusting the noise to a more realistic level. In addition, we will modify the model to generate vectors that can be utilised to construct Toeplitz matrices for off-grid estimates, rather than using a multilabel classification approach.

Data Availability

The data, results, and code for the study in this paper are available from the corresponding author with appropriate attribution.

Conflicts of Interest

We declare that we have no known competing financial interests or personal relationships at the time of submission, and that there are no competing relationships between co-authors.

Authors' Contributions

Wei Wang contributed in the conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing—original draft, writing—review and editing. Lang Zhou contributed in the conceptualisation, formal analysis, investigation, methodology, writing—review and editing. Kun Ye contributed in the conceptualisation, formal analysis, writing—review and editing. Haisin Sun contributed in the conceptualisation, formal analysis, funding acquisition, investigation, methodology, writing—review and editing. Shaohua Hong contributed in the conceptualisation, formal analysis, project administration, supervision, writing—review and editing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62271426).

References

[1] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] X.-L. Xu and K. M. Buckley, "Bias analysis of the MUSIC location estimator," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2559–2569, 1992.

[4] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound: further results and comparisons," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2140–2150, 1990.

[5] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 336–339, IEEE, Boston, MA, USA, April 1983.

[6] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[7] M. Haardt and M. E. Ali-Hackl, "Unitary ESPRIT: how to exploit additional information inherent in the relational invariance structure," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV/229–IV/232, IEEE, Adelaide, SA, Australia, April 1994.

[8] M. Haardt and J. A. Nossek, "Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational burden," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1232–1242, 1995.

[9] C.-L. Liu and P. P. Vaidyanathan, "Remarks on the spatial smoothing step in coarray MUSIC," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1438–1442, 2015.

[10] M. Wagner, Y. Park, and P. Gerstoft, "Gridless DOA estimation and root-MUSIC for non-uniform linear arrays," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2144–2157, 2021.

[11] L. Zhou, K. Ye, J. Qi, and H. Sun, "DOA estimation based on pseudo-noise subspace for relocating enhanced nested array," *IEEE Signal Processing Letters*, vol. 29, pp. 1858–1862, 2022.

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

[13] J.-X. Kou, M. Li, L. Wang, K. Yang, and C.-L. Jiang, "Generalized weight function selection criteria for the compressive sensing based robust DOA estimation methods," *Signal Processing*, vol. 175, Article ID 107663, 2020.

[14] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[15] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[16] Z. Yang and L. Xie, "Enhancing sparsity and resolution via reweighted atomic norm minimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 995–1006, 2016.

[17] B. Liu, S.-y. Matsushita, and L. Xu, "DOA estimation with small snapshots using weighted mixed norm based on spatial filter," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16183–16187, 2020.

[18] H. Wang, X. Wang, X. Lan, T. Su, and L. Wan, "BSBL-based auxiliary vehicle position analysis in smart city using distributed MEC and UAV-deployed IoT," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 975–986, 2023.

[19] J. He, T. Shu, V. Dakulagi, and L. Li, "Simultaneous interference localization and array calibration for robust adaptive beamforming with partly calibrated arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2850–2863, 2021.

[20] T. Shu, J. He, and V. Dakulagi, "3-D near-field source localization using a spatially spread acoustic vector sensor," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 1, pp. 180–188, 2022.

[21] V. Dakulagi, "A new approach to achieve a trade-off between direction-of-arrival estimation performance and computational complexity," *IEEE Communications Letters*, vol. 25, no. 4, pp. 1183–1186, 2021.

[22] V. Dakulagi and J. He, "Improved direction-of-arrival estimation and its implementation for modified symmetric sensor array," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5213–5220, 2021.

[23] J. Cong, X. Wang, X. Lan, and W. Liu, "A generalized noise reconstruction approach for robust DOA estimation," *IEEE Transactions on Radar Systems*, vol. 1, pp. 382–394, 2023.

[24] X. Su, P. Hu, Z. Liu, T. Liu, B. Peng, and X. Li, "Mixed near-field and far-field source localization based on convolution neural networks via symmetric nested array," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7908–7920, 2021.

[25] X. Wu, X. Yang, X. Jia, and F. Tian, "A gridless DOA estimation method based on convolutional neural network with toeplitz prior," *IEEE Signal Processing Letters*, vol. 29, pp. 1247–1251, 2022.

[26] K. SongGong, W. Wang, and H. Chen, "Acoustic source localization in the circular harmonic domain using deep learning architecture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2475–2491, 2022.

[27] H. Xiang, B. Chen, M. Yang, and S. Xu, "Angle separation learning for coherent DOA estimation with deep sparse prior," *IEEE Communications Letters*, vol. 25, no. 2, pp. 465–469, 2021.

[28] Y. Liu, H. Chen, and B. Wang, "DOA estimation based on CNN for underwater acoustic array," *Applied Acoustics*, vol. 172, Article ID 107594, 2021.

[29] G. K. Papageorgiou, M. Sellathurai, and Y. C. Eldar, "Deep networks for direction-of-arrival estimation in low SNR," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3714–3729, 2021.

[30] J. Yu and Y. Wang, "Deep learning-based multipath doAs estimation method for mmWave massive MIMO systems in low SNR," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7480–7490, 2023.

[31] G. Hu, F. Zhao, and B. Liu, "Estimation of the two-dimensional direction of arrival for low-elevation and Non-low-elevation targets based on dilated convolutional networks," *Remote Sensing*, vol. 15, no. 12, Article ID 3117, 2023.

[32] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, and Z. Zhang, "Star-transformer," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1315–1325, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019.

[33] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, Curran Associates Inc., Red Hook, NY, USA, December 2017.

[34] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1783–1795, 1990.

[35] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, Association for Computational Linguistics, New Orleans, Louisiana, June 2018.

[36] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[37] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014.

[38] M. Wang and A. Nehorai, "Coarrays, MUSIC, and the Cramér–Rao bound," *IEEE Transactions on Signal Processing*, vol. 65, no. 4, pp. 933–946, 2017.

[39] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, Wiley, 2002.