

Signal as Token: Robust DOA Estimation in Complex Environments Aided by Transformer

Ziqi Wang^{a,1,*}, Zihan Cao^{a,2,*}, Julian Xie^{a,3,**}, Huiyong Li^a, Zishu He^a

^aUniversity of Electronic Science and Technology of China, Chengdu, 610000, Sichuan Province, China

Abstract

Direction of Arrival (DOA) estimation, with applications in various fields, is a widely-researched problem. However, the lack of adaptation DOA estimation methods in the presence of various complex environments remains a significant challenge in the field. Conventional approaches heavily rely on manually engineered features and assumed array priors. Moreover, their performance is often unsatisfactory in the face of complex environments. In this paper, we propose a novel approach that harnesses the power of the Transformer to tackle the DOA estimation problem. The Transformer leverages the self-attention mechanism to effectively capture long-range dependency within data sequences. By employing the Transformer in the DOA estimation task and introducing an antenna-based attention mechanism tailored for DOA estimation, we provide evidence that the antenna-based attention output corresponds to a pseudo Singular Value Decomposition (pseudo-SVD) of the covariance matrix. Leveraging this mechanism enables us to capture more profound feature information within the received signals, leading to highly accurate DOA estimation. Furthermore, our proposed Transformer-based approach exhibits good adaptability in the presence of a low signal-to-noise (SNR) ratio, a limited number of snapshots, array errors, coherent sources, and broadband sources. Rigorous experiments conducted on synthetic and real-world datasets validate the effectiveness and generalization capability of our method. Additionally, our proposed method has also been proven effective in solving the problem of estimating the number of signal sources. Overall, this work presents a promising solution by seamlessly integrating the capabilities of the Transformer with the DOA estimation task, enabling accurate DOA estimation even in the most challenging scenarios. Our code will be released after possible acceptance.

Keywords: Direction-of-Arrival (DOA) estimation, Transformer, Attention mechanism

1. Introduction

Direction-of-arrival estimation is an extensively researched topic, finding applications in diverse domains such as radar [1], wireless communication [2], UAV localization [3], and sonar detection [4]. With the rapid development of related industries in recent years, DOA estimation has garnered substantial attention. Researchers have diligently focused on developing precise and dependable DOA estimation techniques to cater to the evolving requirements of specific scenes [5].

Classic approaches to Direction of Arrival (DOA) estimation encompass various methods such as array signal processing-based beamforming [6, 7], maximum likelihood estimation [8, 9, 10, 11, 9], subspace-based methods [12, 13, 14, 15], and the

sparsity-inducing methods [16, 17, 18]. These DOA estimation methods critically rely on the intricate statistical properties of the signals. They employ sophisticated signal processing techniques and robust parameter estimation algorithms to infer the precise direction information of the targets. By adeptly utilizing the received signals from the array, these methods effectively exploit the interplay between array geometry and signal statistics, resulting in accurate DOA estimation across diverse application domains.

However, factors such as low SNR, low snapshots, coherent signals, and broadband signals would affect estimation performance seriously. Taking the most classical subspace decomposition method [12, 13, 14] as an example, its core is to decompose the covariance matrix of the received signals into the signal subspace and the noise subspace, relying on the orthogonality between the noise component and the steering vectors for DOA estimation. In order to ensure orthogonality, the above method requires ideal factors (e.g., sufficiently high SNR, many snapshots). If one or more of these factors are not ideal, the performance of general estimation methods typically deteriorates significantly. Scholars have proposed various methods to solve these problems, such as using the MUSIC-like method [19] to deal with the estimation of DOA under low SNR, using the beamforming method [20] to deal with DOA under low snapshots estimation of the problem, utilizing Spatial Smooth-

*Equal contribution to this work.

**Corresponding Author.

¹Z-Q Wang is in the Phased Array and Adaptive Processing Team, School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: ziqiwang327@foxmail.com).

²Z-H Cao is in the Department of Mathematics, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: iamzihan666@gmail.com).

³J-L Xie is a Professor in the Phased Array and Adaptive Processing Team, School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: julanxie@uestc.edu.cn).

ing (SPS) [21] or Compressed Sensing (CS) [22] theory for DOA estimation of correlated signal sources, and employing frequency band division techniques [23] to handle DOA estimation of broadband signal sources. However, these methods often compromise resolution and struggle to accurately recover all sources when the number of sources is unknown.

In addition to the non-ideal factors above, in real-world scenarios, kinds of array errors, such as the inconsistency in amplitude and phase between channels, array element position errors, the mutual coupling, and directional pattern errors of array elements, make formidable challenges in accurately characterizing the array system. These factors undermine the ability to establish precise models, thereby exerting adverse effects on the performance of the aforementioned DOA estimation methods. In order to address this challenge, different array error modeling methods are proposed to estimate DOA [24, 25]. Nevertheless, it is crucial to note that most array error modeling methods rely on stringent assumptions, and any deviation from these assumptions can render the models ineffective. Moreover, in practical scenarios, the presence of multiple concurrent errors can give rise to complex and intertwined effects, which may substantially deviate from the idealized assumptions. These deviations inevitably exert a profound influence on the performance of DOA estimation algorithms, leading to compromised accuracy.

In response to these challenges, researchers have explored the application of machine learning (ML) techniques to address the aforementioned issues. ML-based approaches, such as Support Vector Machines (SVM) [26, 27], Support Vector Regression (SVR) [28, 29], and Radial Basis Function (RBF) [30, 31], have emerged as promising solutions, showcasing enhanced modeling capabilities. These data-driven methods operate by leveraging training data, where input data and corresponding angle labels are used to establish a non-linear mapping from inputs to outputs. Notably, these methods exhibit a distinct advantage by not relying on assumptions regarding array geometry, thus demonstrating robustness in the face of non-ideal factors. In fact, researches [29] have substantiated their superiority over traditional subspace-based methods (e.g., MUSIC) in terms of effectiveness and performance in the DOA estimation task.

Nonetheless, ML-based methods still exhibit notable limitations. Although they showcase commendable adaptability to array geometries, the efficacy of these estimation approaches primarily hinges upon the learning proficiency and generalization prowess of models. A pressing concern lies in the fact that the majority of DOA estimation works employing machine learning techniques predominantly rely on synthetic data, which may diverge from the complexities inherent in real-world scenarios. Consequently, such disparities can engender a decline in performance for these methods when confronted with practical applications.

In recent years, there has been rapid development and widespread application of deep learning-based (DL-based) methods in various fields, such as image classification, image generation, and sequential modeling. These methods leverage deep neural network architectures, including Convolutional Neural Networks (CNNs) [32, 33, 34], Recurrent Neural Networks (RNNs) [35, 36], and Transformers [37, 38, 39, 40, 41],

to learn from data and make accurate predictions. DL-based methods exhibit stronger capabilities in fitting complex mappings and extracting deeper features compared to traditional machine-learning approaches. Given the advantages of deep learning, researchers have also explored its potential in solving the DOA estimation problem. Notably, Liu *et al.* [42] introduced the use of an autoencoder to address the challenge of subspace partitioning. An autoencoder is employed to divide the input covariance matrix into multiple angle grids, enabling coarse classification of the DOA on the grid for each source. Fine-grained angle estimation within each grid was achieved using a Multilayer Perceptron (MLP), and the final results were obtained through bilinear interpolation of the predictions of the MLP. It demonstrated promising performance in general DOA estimation tasks and showed some adaptability to array errors. Nonetheless, it not only needs tedious two-stage training and faces challenges in scenarios with low SNR ratios and coherent sources. Furthermore, the heavy reliance on classifiers in the overall model architecture limited its performance due to grid constraints. Shmuel *et al.* [43] utilized a Deep Convolutional Neural Network (DCNN) to reconstruct the covariance matrix of received signals. This reconstructed matrix was then integrated into traditional subspace-based DOA estimation methods, combining the interpretability of traditional methods with the powerful learning capabilities of deep models. However, it should be noted that the output matrix lacks explicit constraints, and the model can only be trained with differentiable traditional methods.

Based on the above, previous DL-based DOA estimation works have often been limited to MLP or shallow CNN, overlooking the impact of deep models on DOA estimation. Furthermore, most previous works have restricted DOA estimation to angle grids, transforming DOA estimation into a classification problem on the grid. Although these approaches reduce the learning difficulty for the network, they can lead to inaccurate DOA estimation. Transformer was initially introduced in the field of natural language processing (NLP) [44, 41] and has since been applied to natural images [45], medical images [46, 47], sequential modeling [48], and content generation [37, 38]. Importantly, the attention mechanism is the key component of the Transformer, calculating an element-wise correlation matrix thereby assigning different weights to different elements. As a result, the attention mechanism allows the model to focus on the parts of the data that are crucial for the task. Therefore, incorporating the Transformer architecture into DOA estimation has tremendous potential.

Note that, the obtaining of the traditional attention map is computed by covariance-like matrix multiplication on the channel dimension, thereby resulting in snapshot-length square attention. However, the covariance matrix generated by signals is proven to be effective in extracting DOA features by most model-based methods but is in conflict with the traditional snapshot-length square attention. Inspired by this finding, we propose a novel attention mechanism that models the correlation of antennas and outputs antenna-length square attention. By utilizing the attention mechanism, the proposed model can capture complex relationships and dependencies, while achiev-

ing more accurate and robust DOA estimation. Additionally, the flexibility of the proposed attention mechanism enables the model to adapt to various data patterns and effectively handle challenging scenarios.

The contributions of this paper can be summarized in the following four folds

1. Instead of the DL-based conventional classification method, we change the modeling objective into direct DOA regression. This modification allows the network to estimate DOA more accurately and minimize errors.
2. Furthermore, we enhance the original Transformer architecture by incorporating the proposed effective attention mechanism. This modification enables the model to learn the effective feature representations from the covariance matrix. Moreover, the proposed attention mechanism has better interpretability.
3. Extensive experiments on the DOA task including low-SNR, limited snapshots, coherent, broadband sources, and array errors show that our proposed method can obtain the least DOA estimation error and is robust to these non-ideal factors.
4. We validate the effectiveness of our proposed model on real-world datasets, which enhances the practicability of our model.

The rest of the paper is organized as follows: in Sec. 2, we present the signal model and traditional attention mechanism. In Sec. 3, we introduced the proposed antenna-based attention mechanism. In Sec. 4, we present simulation results in various situations. In Sec. 5, we summarize and highlight our conclusions.

2. Preliminary

2.1. Conventional Signal Modeling

Assuming K signals with the center frequency of f impinge on a uniform line array composed of M omni-directional antennas, with inter-antenna spacing $d = \frac{\lambda}{2}$, where λ represents the wavelength, and their DOAs denoted as $\theta_1, \dots, \theta_K$. The k -th source signal can be represented as $s_k(t)$. The array output is obtained by uniformly sampling the received signal $\mathbf{x}(t)$ at N time instants, resulting in a total of N snapshots. The received vector can be represented as follows: $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(N)]$. The array outputs are contaminated by the zero-mean Gaussian noise $\mathbf{v}(n)$.

In the absence of array errors, the mapping from the wave direction to the array output should be accurately established. This means that there exists a relationship between the array output and the signal input, which can be expressed as follows,

$$\mathbf{x}(n) = \sum_{k=1}^K \mathbf{a}(\theta_k) s_k(n) + \mathbf{v}(n), n = 1, \dots, N \quad (1)$$

Its matrix form can be expressed as follows,

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{V}, \quad (2)$$

where,

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_M(1) & x_M(2) & \cdots & x_M(N) \end{bmatrix} \in \mathbb{C}^{M \times N}, \quad (3)$$

$$\mathbf{S} = \begin{bmatrix} s_1(1) & s_1(2) & \cdots & s_1(N) \\ s_2(1) & s_2(2) & \cdots & s_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ s_K(1) & s_K(2) & \cdots & s_K(N) \end{bmatrix} \in \mathbb{C}^{K \times N}, \quad (4)$$

$$\mathbf{V} = \begin{bmatrix} v_1(1) & v_1(2) & \cdots & v_1(N) \\ v_2(1) & v_2(2) & \cdots & v_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ v_M(1) & v_M(2) & \cdots & v_M(N) \end{bmatrix} \in \mathbb{C}^{M \times N}, \quad (5)$$

$$\mathbf{A} = [\mathbf{a}(\theta_1) \quad \mathbf{a}(\theta_2) \quad \cdots \quad \mathbf{a}(\theta_K)] \in \mathbb{C}^{M \times K}, \quad (6)$$

where $x_i(j)$ represents the j -th snapshot received by the i -th array element, $s_i(j)$ represents the j -th snapshot of the i -th source signal, and $v_i(j)$ represents the j -th snapshot noise of the i -th array element.

However, practical arrays suffer from various errors, such as the inconsistency in amplitude and phase between channels, array element position errors, the mutual coupling, and directional pattern errors of array elements. The presence of these errors introduces the steering vector $\mathbf{a}(\theta)$ mismatch, rendering the ideal assumptions invalid. The amplitude and phase inconsistency between channels is generally unrelated to the DOA of the signal. Thus, the array output with this type of error can be modeled as,

$$\mathbf{x}(n) = [\mathbf{L}\mathbf{A}]\mathbf{s}(n) + \mathbf{v}(n), \quad (7)$$

where,

$$\mathbf{L} = \text{diag}[l_1, l_2, \dots, l_m, \dots, l_M] \in \mathbb{C}^{M \times M}, \quad (8)$$

where l_m is a complex number that represents amplitude and phase errors for the m -th antenna. The array element position errors will introduce a phase error with directional dependence. Therefore, the observed data of the array can be modeled as,

$$\mathbf{x}(n) = [\mathbf{A} \odot \mathbf{B}]\mathbf{s}(n) + \mathbf{v}(n), \quad (9)$$

where \odot represents the Hadamard product.

$$\mathbf{B} = [\mathbf{b}(\theta_1) \quad \mathbf{b}(\theta_2) \quad \cdots \quad \mathbf{b}(\theta_K)] \in \mathbb{C}^{M \times K}, \quad (10)$$

is the matrix representing the errors introduced by array element positions, where the error vector can be represented as,

$$\mathbf{b}(\theta_k) = [e^{j\Delta\phi_1(\theta_k)}, \dots, e^{j\Delta\phi_M(\theta_k)}]^T, \quad (11)$$

$\Delta\phi_m(\theta_k)$ is the phase error introduced by the relative position of the m -th element with respect to the k -th impinge direction. The mutual coupling between the array elements enables the array output to be represented as,

$$\mathbf{x}(n) = \mathbf{C}\mathbf{A}\mathbf{s}(n) + \mathbf{v}(n). \quad (12)$$

where \mathbf{C} is the mutual coupling matrix,

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1M} \\ c_{21} & c_{22} & \cdots & c_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \cdots & c_{MM} \end{bmatrix} \in \mathbb{C}^{M \times M}. \quad (13)$$

The directional pattern error mainly refers to the gain errors between the actual directional pattern and the ideal directional pattern. The errors it introduces also have directional dependence. The array output considering this errors can be represented as,

$$\mathbf{x}(n) = [\mathbf{A} \odot \mathbf{P}]\mathbf{s}(n) + \mathbf{v}(n), \quad (14)$$

where,

$$\mathbf{P} = [\mathbf{p}(\theta_1) \quad \mathbf{p}(\theta_2) \quad \cdots \quad \mathbf{p}(\theta_K)] \in \mathbb{C}^{M \times K}, \quad (15)$$

$$\mathbf{p}(\theta) = [p_0(\theta), \cdots, p_{M-1}(\theta)]^T. \quad (16)$$

When all the aforementioned non-ideal factors are presented, the receiving model of the array can be expressed as,

$$\mathbf{x}(n) = \{\mathbf{L}[\mathbf{C}(\mathbf{A} \odot \mathbf{P} \odot \mathbf{B})]\mathbf{s}(n) + \mathbf{v}(n). \quad (17)$$

In this case, the steering matrix of the array, also known as the array response matrix, is modified to be,

$$\tilde{\mathbf{A}} = \mathbf{L}[\mathbf{C}(\mathbf{A} \odot \mathbf{P} \odot \mathbf{B})]. \quad (18)$$

Therefore, we obtain the following equations,

$$\mathbf{x}(n) = \tilde{\mathbf{A}}\mathbf{s}(n) + \mathbf{v}(n), \quad (19)$$

$$\tilde{\mathbf{a}}(\theta) = \mathbf{L}[\mathbf{C}(\mathbf{a}(\theta) \odot \mathbf{p}(\theta) \odot \mathbf{b}(\theta))]. \quad (20)$$

It can be clearly observed that the error in the steering vector is frequency and direction-dependent. Despite the modeling approaches mentioned above, it is often challenging to accurately model for complex real-world environments. Thus, approximation and simplification methods to the data modeling are commonly employed in many DOA estimation methods. They face the problem of a sharp decline in DOA estimation performance once approximation and simplification fail. In addition to array errors affecting the accuracy of DOA estimation, the correlation between signals and broadband signals mentioned earlier can also impact the estimation performance. Traditional methods designed for correlated or broadband signals often sacrifice resolution, which can affect the estimation accuracy and may not meet the specific requirements of practical scenarios. Therefore, it is crucial to explore a framework or approach that can effectively address these problems.

2.2. Traditional Attention

We first review the traditional attention mechanism. The attention mechanism can be seen as a data-adaptive weighted operation. It extracts different parts weighted by the covariance-like attention map on the spatial dimension, which can be formulated as follows,

$$\mathcal{A}_S = \mathbf{Q}^T \mathbf{K}, \quad (21)$$

where $\mathcal{A}_S \in \mathbb{R}^{N \times N}$ is the covariance-like attention map, M is the number of feature channels. $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M \times N}$ are named query, key, value and defined as follows,

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{W}^Q \mathbf{g}(\mathbf{X}), \mathbf{W}^K \mathbf{X}, \mathbf{W}^V \mathbf{X}. \quad (22)$$

The $\mathbf{W}^i (i = Q, K, V)$ is a linear projection weight matrix⁴, which is responsible for projecting the input feature $\mathbf{X} \in \mathbb{R}^{M \times N}$ into three different high-dimensional spaces. g is a matrix block transformation. When g is a specific transformation rather than an identity transformation (i.e., $\mathbf{X} = g(\mathbf{X})$), the attention is a cross-attention otherwise is a self-attention. Then the output of the weighted attention operation is defined as follows,

$$\mathbf{O} = \text{Softmax}\left(\frac{\mathcal{A}_S}{\sqrt{M}}\right) \mathbf{V}^T, \quad (23)$$

where Softmax is applied to ensure the attention map is probabilistic in each row. This equation shows that the information in \mathbf{V} space is projected into \mathbf{O} space by attention map \mathcal{A}_S .

In the context of DOA estimation, the attention map $\mathcal{A}_S \in \mathbb{R}^{N \times N}$ is obtained by a matrix multiplication by two projected matrixs. It models the relationship within each snapshot, then results in $N \times N$ attention map.

3. Proposed System Model

3.1. Math Notation

In this section, we elaborate on the overall math notion represented in Tab. 1.

3.2. Overall Architecture

We first introduce the overall architecture illustrated in Fig. 1. Our DOA estimation deep architecture is a pure Transformer whose input signals \mathbf{X} are shaped as $\mathbb{C}^{M \times N}$. The Transformer architecture is formed by L Transformer layers. The first layer takes the tokenized signal tokens $\mathbf{X} \in \mathbb{R}^{M \times 2N}$ as input and processes the tokens to feed into the next layer. It is worth noting that we retain the term ‘‘tokenization’’ here, as it was introduced in the original Transformer paper [44] and has been used since then. In practice, we treat each signal snapshot as a token. The tokens \mathbf{X} are defined as follows,

$$\mathbf{X} = [\mathbf{R}(\mathbf{X}) \ \mathbf{I}(\mathbf{X})], \quad (24)$$

where $[\cdot]$ is concatenating along the last dimension. We will elaborate on the rationale behind this concatenation in Section 3.3. Before feeding the first layer, we add a learnable token $[\text{reg}] \in \mathbb{R}^{M \times 2}$ into the original tokens for regression DOA.

⁴For simplicity, we omit the bias part here.

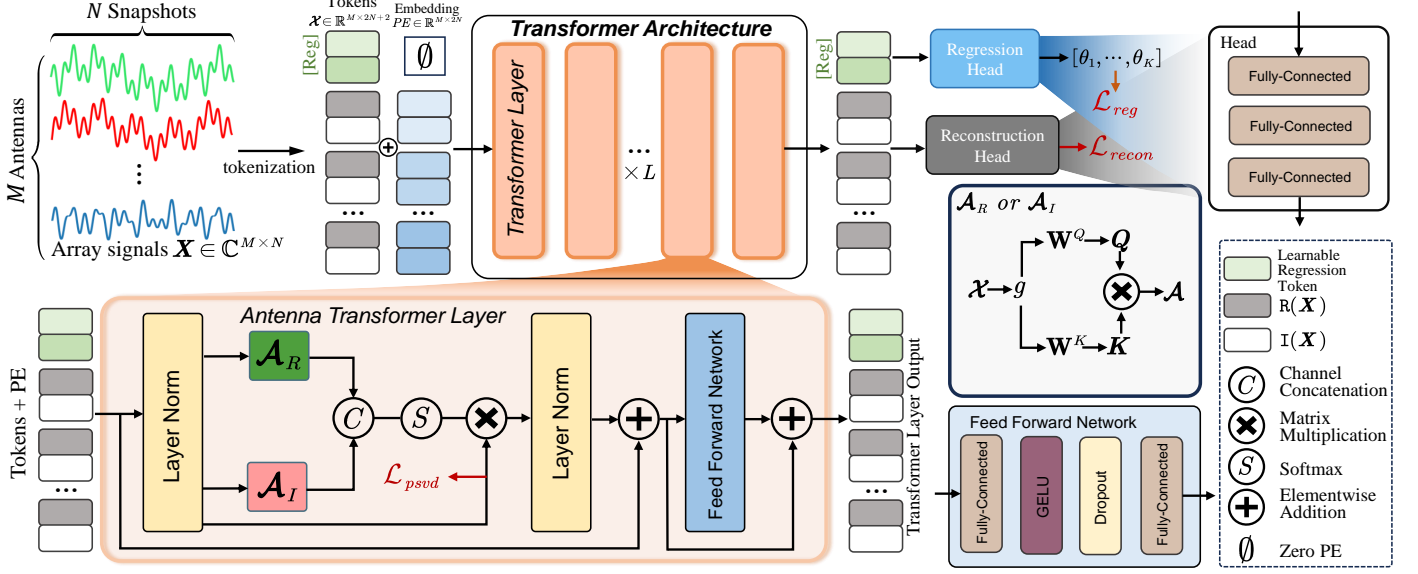


Figure 1: Overview of the proposed model architecture. **Upper panel:** The original signal sources are first tokenized into tokens \mathbf{X} and stacked with the real and imaginary parts on the antenna dimension. The tokens are concatenated with a [reg] token and added elementwisely with positional embedding (i.e., PE) for final regression and are fed into the L antenna Transformer layers. Then the [reg] token is extracted to a regression head to output the final DOA. **Left bottom panel:** Illustration of our proposed antenna-based attention module. **Right bottom panel:** the detailed illustration of the proposed effective antenna-based attention (see Sect. 3.3) and the composition of the feed forward network.

Table 1: Math notation.

Notation	Explanation
\mathbf{X}	Received signals.
\mathbf{X}	The input of the Transformer layer (i.e., tokens).
\mathbf{R}	Covariance matrix obtained by Eq. 31.
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, Key, Value defined in Eq. 22.
$\mathcal{A}_S, \mathcal{A}_R, \mathcal{A}_I$	Attention maps formed in Eqs. 21, 32.
\mathbf{O}	Output features from attention defined in Eq. 23.
$(\cdot)^H$	Matrix conjugate transpose.
$(\cdot)^*$	Matrix conjugate.
$(\cdot)^T$	Matrix transpose.
K	Number of sources.
M	Number of antennas.
N	Number of snapshots.
Z	Number of samples.
$\text{R}(\cdot)$	The real part of a complex-valued matrix.
$\text{I}(\cdot)$	The imaginary part of a complex-valued matrix.
$f(\cdot)$	The proposed Transformer model.

To make the Transformer model sensitive to positional information, we add positional encoding (PE) $\in \mathbb{R}^{1 \times 2N}$ to the array token directly. PE can be expressed as the following formula,

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/M}), \quad (25a)$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/M}). \quad (25b)$$

This gives a positional encoding of shape $1 \times N$. We repeat this positional encoding twice to obtain a positional encoding of shape $1 \times 2N$ (since each two consecutive tokens are composed of the real and imaginary parts of a snapshot, they should share the same positional encoding). Note that, the regression learnable tokens do not add with positional encoding.

After L layers, the output of our proposed model is shaped as $\mathbb{R}^{M \times (2N+2)}$. We take out the regression DOA token and send it to the regression head, which is a simple MLP composed of three-layer Fully-Connected (FC) layers, to get an estimated DOA. The rest of the $2N$ tokens are projected into the tokenized signal space to reconstruct the original tokens. The aim of the reconstruction is to constrain the attention map in high-dimensional space (see Corollary. 3.1).

The detailed design of each Transformer layer is shown in the left bottom panel of Fig. 1. Firstly, the input tokens are fed into a LayerNorm, which operates normalization on channel dimension (i.e., M dimension in the DOA context). Then two outcomes of $\mathcal{A}_R, \mathcal{A}_I$ attention mechanisms are possessed and concatenated together into another LayerNorm. To facilitate training stability, a shortcut is enabled, which adds the input to the normalized intermediate features. After that, the features are fed into a Feed Forward Network (FFN) to enhance the feature representation. The FFN is implemented by an MLP with two fully-connected (FC) layers. The first FC layer doubles the number of channels and the last FC layers reduce the number of channels back to M . A GELU activation [49] is inserted between the first and the second FC layer. Similarly, another shortcut is added. The overall pipeline can be formulated as follows,

$$\mathbf{X}_{norm}^l = \text{Norm}(\mathbf{X}^l), \quad (26a)$$

$$\mathbf{X}_{attm}^l = \text{Softmax}\left(\frac{[\mathcal{A}_R \mathcal{A}_I]^T}{\sqrt{2M}}\right) \mathbf{V}, \quad (26b)$$

$$\mathbf{X}_{norm2}^l = \text{Norm}(\mathbf{X}_{attm}^l) + \mathbf{X}^l, \quad (26c)$$

$$\mathbf{X}_{FFN}^l = \text{FFN}(\mathbf{X}_{norm2}^l), \quad (26d)$$

$$\mathbf{X}^{l+1} = \mathbf{X}_{FFN}^l + \mathbf{X}_{norm2}^l, \quad (26e)$$

where superscript l denotes the l^{th} layer of the Transformer network. The training objective of our model is regression rather than classification in [42] nor the rectified covariance matrix \mathcal{R} in [43]. Our training objective contains three terms: regression, reconstruction, and regularization objectives, which can be formulated as corresponding three loss functions,

$$\mathcal{L}_{reg} = \|\text{MLP}_{reg}(f(\mathcal{X})), \text{DOA}\|_2^2, \quad (27)$$

$$\mathcal{L}_{recon} = \|\text{MLP}_{recon}(f(\mathcal{X})), \mathcal{X}\|_2^2, \quad (28)$$

$$\mathcal{L}_{psvd} = \|\mathbf{V}^T \mathbf{V} - 2\mathbf{E}\|_2^2 + \|\mathbf{V} \mathbf{V}^T - 2\mathbf{E}\|_2^2. \quad (29)$$

The MLP_{reg} , MLP_{recon} are regression head and reconstruction head, respectively. DOA is used as the label for supervising the training. \mathcal{L}_{psvd} is a regularization loss which regularizes the attention output can be pseudo singular value decomposition (see more details in Theorem 3.1). In the training phase, we optimize the combination of the three losses together which is denoted as follows,

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \mathcal{L}_{recon} + 0.1\mathcal{L}_{psvd}. \quad (30)$$

3.3. Effective antenna-based Attention

Due to the success of many model-based methods, they often form the covariance matrix⁵,

$$\mathcal{R} = \mathcal{X}\mathcal{X}^H, \quad (31)$$

and process it to extract DOA-related features (e.g., decomposing into noise and signal subspaces). Note the similarity between the commonly used covariance matrix and the attention map, it is easy to find that the covariance matrix links the relations within antennas rather than snapshots brought by the attention map. We speculate that the traditional attention that models relationships among snapshots is insufficient and ineffective.

We propose a novel antenna-based covariance-like attention mechanism to extract features from the well-studied signal covariance matrix \mathcal{R} . The attention map is divided into two terms (i.e., $\mathcal{A}_R, \mathcal{A}_I$) to model the real part and imaginary part of the covariance matrix. The division formula is formed as follows,

$$\mathcal{A} := [\mathcal{A}_R \mathcal{A}_I] \in \mathbb{R}^{M \times 2M}. \quad (32)$$

\mathcal{A} is a tensor concatenated by \mathcal{A}_R and \mathcal{A}_I , which means the real and imaginary parts are stacked on the last dimension. With this division, we can model the covariance matrix \mathcal{R} in attention. The traditional attention can be reformulated into our antenna-

based attention as follows,

$$\mathcal{Q}_R, \mathcal{K}_R = \mathbf{W}_R^Q g(\mathcal{X}), \mathbf{W}_R^K \mathcal{X}, \quad (33a)$$

$$\mathcal{Q}_I, \mathcal{K}_I = \mathbf{W}_I^Q g(\mathcal{X}), \mathbf{W}_I^K \mathcal{X}, \quad (33b)$$

$$\mathcal{V} = \mathbf{W}^V \mathcal{X}, \quad (33c)$$

$$\mathcal{A}_R = \mathcal{Q}_R \mathcal{K}_R^T, \quad (33d)$$

$$\mathcal{A}_I = \mathcal{Q}_I \mathcal{K}_I^T \quad (33e)$$

$$\mathcal{A} = [\mathcal{A}_R \mathcal{A}_I], \quad (33f)$$

$$\mathcal{O} = \text{Softmax}\left(\frac{\mathcal{A}^T}{\sqrt{2M}}\right) \mathcal{V}, \quad (33g)$$

where \mathcal{X} is the input of the module from the previous Transformer layer⁶, g is a matrix block transformation, and $\text{Softmax}(\mathbf{A}_{i,j}) = \exp(\mathbf{A}_{i,j}) / \sum_i \exp(\mathbf{A}_{i,j})$. $\mathbf{W}_j^i, i = (Q, K, V), j = (R, I, \emptyset)$ are projection weights that project the inputs into a high-dimensional space. Based on the calculation of $\mathcal{A}_R, \mathcal{A}_I$, the input of the first Transformer layer is the tokenized signal that the real and imaginary parts are stacked along the snapshot dimension (i.e., $[\mathcal{R}(\mathcal{X}) \ \mathcal{I}(\mathcal{X})]$). Specifically, taking the first transformer layer as an example, when computing \mathcal{A}_R , g is the identity transformation,

$$g(\mathcal{X}) = \mathcal{X} = [\mathcal{R}(\mathcal{X}) \ \mathcal{I}(\mathcal{X})]. \quad (34)$$

When computing \mathcal{A}_I ,

$$g(\mathcal{X}) = [\mathcal{X}_{(N+1:2N)} \ -\mathcal{X}_{(1:N)}], \quad (35)$$

where $(m : n)$ are indexing operation from index m to n . So our antenna-based attention is hybrid attention (i.e., self-attention and cross-attention). According to the above, the covariance matrix can be modeled in our antenna-based attention,

$$\mathcal{R} = \underbrace{[\mathcal{R}(\mathcal{X}) \ \mathcal{I}(\mathcal{X})] \cdot [\mathcal{R}(\mathcal{X})^T \ \mathcal{I}(\mathcal{X})^T]}_{\mathcal{A}_R} + j \underbrace{[\mathcal{I}(\mathcal{X}) \ -\mathcal{R}(\mathcal{X})] \cdot [\mathcal{R}(\mathcal{X})^T \ \mathcal{I}(\mathcal{X})^T]}_{\mathcal{A}_I}. \quad (36)$$

We present the remarks and prove the equality of the produced attention map \mathcal{A} and the stacked covariance matrix $[\mathcal{R}(\mathcal{R}) \ \mathcal{I}(\mathcal{R})]$.

Remark 3.1 (Covariance real part equality). *Given a complex-valued signal input \mathcal{X} , the antenna-based attention first term \mathcal{A}_R equals the real part $\mathcal{R}(\mathcal{R})$ of the covariance matrix for the complex-valued signal.*

Proof 3.1. *Equality can be proven by using the property of complex-valued matrix: $\mathcal{R}(\mathcal{X}^H) = \mathcal{R}(\mathcal{X})^T, \mathcal{I}(\mathcal{X}^H) = -\mathcal{I}(\mathcal{X})^T$. Then, the \mathcal{A}_R can be derived by following equations,*

$$\mathcal{A}_R = \mathcal{R}(\mathcal{X}) \cdot \mathcal{R}(\mathcal{X})^T + \mathcal{I}(\mathcal{X}) \cdot \mathcal{I}(\mathcal{X})^T \quad (37a)$$

$$= \mathcal{R}(\mathcal{X}) \cdot \mathcal{R}(\mathcal{X}^H) - \mathcal{I}(\mathcal{X}) \cdot \mathcal{I}(\mathcal{X}^H) \quad (37b)$$

$$= \mathcal{R}(\mathcal{X}\mathcal{X}^H) = \mathcal{R}(\mathcal{R}). \quad (37c)$$

⁵For the sake of brevity, we have omitted the denominator N , which is used to divide the variable \mathcal{R} .

⁶We omit the superscript l for clear expression.

Remark 3.2 (Covariance imaginary part equality). Given a complex-valued signal input \mathbf{X} , the antenna-based attention second term \mathcal{A}_I equals the imaginary part $\mathbf{I}(\mathcal{R})$ of the covariance matrix for the complex-valued signal.

Proof 3.2. Similar to Proof. 3.1, equality can be proven by using the property of complex-valued matrix, then the \mathcal{A}_I can be derived by following equations,

$$\mathcal{A}_I = -\mathbf{R}(\mathbf{X}) \cdot \mathbf{I}(\mathbf{X})^T + \mathbf{I}(\mathbf{X}) \cdot \mathbf{R}(\mathbf{X})^T \quad (38a)$$

$$= \mathbf{R}(\mathbf{X}) \cdot \mathbf{I}(\mathbf{X}^H) + \mathbf{I}(\mathbf{X}) \cdot \mathbf{R}(\mathbf{X}^H) \quad (38b)$$

$$= \mathbf{I}(\mathbf{X}\mathbf{X}^H) = \mathbf{I}(\mathcal{R}). \quad (38c)$$

We find that our antenna-based attention can be pseudo singular value decomposition, which enhances interpretability.

Theorem 3.1 (Antenna-based attention pseudo-SVD theorem).

Considering a well-trained Transformer model, the antenna-based attention \mathcal{A} can approximate the stacked covariance matrix $[\mathbf{R}(\mathcal{R}) \ \mathbf{I}(\mathcal{R})]$. The output of the attention module \mathbf{O} can be singular value decomposition (SVD) into $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = \left(\frac{\mathbf{T}^H\mathbf{Q}^*}{\sqrt{2}}\right)\mathbf{\Sigma}\left(\frac{\mathbf{Q}^T\mathbf{V}}{\sqrt{2}}\right)^H$, where the \mathbf{Q} and $\mathbf{\Sigma}$ are the eigenvectors and eigenvalues of \mathcal{R} , respectively. $\mathbf{T} = [\mathbf{E} \ \mathbf{J}]$, where \mathbf{E} is the unit matrix and $\mathbf{J} = j \cdot \mathbf{E}$.

It means that the output of the proposed attention module is in the space spanned by the (projected) signal and noise eigenvectors. Furthermore, the eigenvalues in $\mathbf{\Sigma}$ define the importance of the eigenvectors. Akin to the MUSIC algorithm, if the eigenvalue is large, the spanned space is dominated by the corresponding eigenvector, which forces the model to concentrate on the vital eigenvectors. We prove this theorem as follows.

Proof 3.3. If the Transformer is trained converged, the output of the attention module is obtained and defined as,

$$\mathbf{O} = \mathcal{A}^T \mathbf{V} = [\mathbf{R}(\mathcal{R}) \ \mathbf{I}(\mathcal{R})]^T \mathbf{V} = \begin{bmatrix} \mathbf{R}(\mathcal{R})^T \mathbf{V} \\ \mathbf{I}(\mathcal{R})^T \mathbf{V} \end{bmatrix} \quad (39)$$

Then we define the pseudo-complex-value output of \mathbf{O} as,

$$\widehat{\mathbf{O}} = \mathbf{R}(\widehat{\mathbf{O}}) + j\mathbf{I}(\widehat{\mathbf{O}}) = \mathbf{T}\mathbf{O}. \quad (40)$$

Then we can get,

$$\mathbf{R}(\widehat{\mathbf{O}}) = \mathbf{R}(\mathcal{R})^T \mathbf{V}, \quad (41a)$$

$$\mathbf{I}(\widehat{\mathbf{O}}) = \mathbf{I}(\mathcal{R})^T \mathbf{V}, \quad (41b)$$

The covariance matrix can be eigendecomposed as follows,

$$\mathcal{R} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^H. \quad (42)$$

We can reformulate the output into,

$$\widehat{\mathbf{O}} = \mathbf{R}(\mathcal{R})^T \mathbf{V} + j\mathbf{I}(\mathcal{R})^T \mathbf{V} \quad (43a)$$

$$= (\mathbf{R}(\mathcal{R})^T + j\mathbf{I}(\mathcal{R})^T) \mathbf{V} \quad (43b)$$

$$= \mathcal{R}^T \mathbf{V} \quad (43c)$$

$$= (\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^H)^T \mathbf{V}. \quad (43d)$$

Refer to Eq. 40, equality can be derived as follows,

$$(\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^H)^T \mathbf{V} = \mathbf{T}\mathbf{O} \quad (44a)$$

$$\Rightarrow \mathbf{O} = \mathbf{T}^\dagger \mathbf{Q}^* \mathbf{\Sigma} \mathbf{Q}^T \mathbf{V} \quad (44b)$$

$$= \frac{1}{2} \mathbf{T}^H \mathbf{Q}^* \mathbf{\Sigma} \mathbf{Q}^T \mathbf{V}. \quad (44c)$$

We prove this by considering $\frac{\mathbf{T}^H \mathbf{Q}^*}{\sqrt{2}}$ and $\frac{\mathbf{Q}^T \mathbf{V}^H}{\sqrt{2}}$ as the unitary matrixs \mathbf{U} and \mathbf{V} , respectively. The new unitary matrix still holds the orthogonality that can be proven following,

$$\mathbf{U}\mathbf{U}^H = \frac{1}{2} \mathbf{T}^H \mathbf{Q}^* \mathbf{Q}^T \mathbf{T} = \frac{1}{2} \mathbf{T}^H (\mathbf{Q}\mathbf{Q}^H)^* \mathbf{T} \quad (45a)$$

$$= \frac{1}{2} \mathbf{T}^H \mathbf{T} = \mathbf{E} \quad (45b)$$

$$\mathbf{U}^H \mathbf{U} = \frac{1}{2} \mathbf{Q}^T \mathbf{T} \mathbf{T}^H \mathbf{Q}^* = (\mathbf{Q}^T \mathbf{Q})^* = (\mathbf{Q}^H \mathbf{Q})^* \quad (45c)$$

$$= \mathbf{E}. \quad (45d)$$

Consequently, to ensure that the \mathbf{V} matrix holds the orthogonality, we can add another regularization term $\|\mathbf{V}^T \mathbf{V} - 2\mathbf{E}\|_2^2$. The rationale can be proven as follows,

$$\mathbf{V}\mathbf{V}^H = \frac{1}{2} \mathbf{V}^H \mathbf{Q}^* \mathbf{Q}^T \mathbf{V} = \frac{1}{2} \mathbf{V}^T (\mathbf{Q}\mathbf{Q}^H)^* \mathbf{V} \quad (46a)$$

$$= \frac{1}{2} \mathbf{V}^T \mathbf{V}, \quad (46b)$$

$$\mathbf{V}^H \mathbf{V} = \frac{1}{2} \mathbf{Q}^T \mathbf{V} \mathbf{V}^T \mathbf{Q}^*. \quad (46c)$$

Similarly, $\frac{1}{2} \mathbf{V}^T \mathbf{V}$ and $\frac{1}{2} \mathbf{V} \mathbf{V}^T$ should all be unit matrixs. As presented in Eq. 29, after performing regularization loss, the matrix \mathbf{V} can hold the orthogonality.

Another by-product of our antenna-based attention is the reduction of memory. It can be seen that the memory complexity of the traditional attention is $O(N^2)$, so when the number of snapshots is large (e.g. usually 1000 or even larger) or the model is deep, the memory consumption is unaffordable. While our antenna-based attention is $O(2M^2) \approx O(M^2)$ memory complexity, it is more device friendly.

3.4. Antenna-based Attention on High Dimension

According to the theory of information bottleneck [50, 51], the model tends to drop out useless information for the training objective, which naturally forms an information bottleneck. In our context, the attention maps \mathcal{A} are in high-dimensional space which is not bounded. In other words, the high-dimensional \mathbf{X} may suffer information loss when the training objective is just the DOA regression.

Corollary 3.1 (High-dimension attention map constrain).

The attention maps $\mathcal{A}_R, \mathcal{A}_I$ are bounded to obtain the lossless information of covariance matrix \mathcal{R} on high-dimensional space when adding a reconstruction loss on the output.

Performing reconstruction loss on the output of the model regardless of [reg] token can force the model to maintain the information and the attention map \mathcal{A} can be seen as a high-dimensional projected version of \mathcal{R} . In this way, the attention maps are constrained.

4. Experiments

4.1. Datasets

We train and test the proposed model using simulated data, which was generated with multiple adjustable parameters including SNR, number of signal sources (K), number of snapshots (N), coherence, and broadband characteristics. We set the number of array elements $M = 10$, inter-antenna spacing $d = \frac{\lambda}{2}$, frequency $f = 200$ MHz (only for narrowband signal), and the range of impinge angles from -60 to 60 degrees. For multi-source cases, the DOA of each source is randomly selected within the impinge angle range. Our basic setting for variable parameters is defined as SNR = 0 dB, $K = 2$, $N = 100$, narrowband signals, non-coherent sources, and without array errors. When generating the data, we vary one variable parameter at each time while keeping the others fixed to explore the adaptability to different parameters of the model.

It should be noted that according to the definition of SNR in Eq. 47, for a received signal containing K sources, we select the SNR calculated from the maximum power source. This means that our SNR is actually an upper limit measurement, in which the SNR corresponding to any other source will not exceed the SNR value in Eq. 47. In other words, the average SNR of the entire received signal is generally lower than the result of Eq. 47. In the experiment, we controlled the difference between the SNR of each source and the SNR in Eq. 47 to not exceed 3 dB.

Subsequently, we examine complex scenarios where multiple variable parameters are changed simultaneously. Each scenario above consists of a training set with 10000 samples and a testing set with 2000 samples. Each sample is composed of an array-received signal \mathbf{X} and the corresponding DOA, where \mathbf{X} serves as the input to the model and DOA is used as the groundtruth. After training and testing the model with simulated data, we further evaluate its applicability by testing it with real-world data, which consists of single-frequency signals with a frequency of 30 KHz, snapshots $N = 512$, SNR = 5 dB, number of array elements $M = 5$, and a uniform circular array. There are 20 samples in this real-world dataset in total.

4.2. Benchmarking

We reimplement several model-based methods and recent state-of-the-art (SOTA) DL-based methods for comprehensive comparisons. The model-based methods are tested with meticulous parameter tuning, while the DL-based methods are trained with full convergence. The model-based methods contain MUSIC [52], ESPRIT [12], Root-MUSIC (R-MUSIC) [53], Spatial Smoothing-MUSIC (SPS-MUSIC) [54], Spatial Smoothing-Root-MUSIC (SPS-R-MUSIC) [54], Spatial Smoothing-ESPRIT (SPS-ESPRIT) [54], BroadBand-MUSIC (BB-MUSIC) [23] and BroadBand-ESPRIT (BB-ESPRIT) [55]. The DL-based methods include SubspaceNet [43] and DOA-Autoencoder [42].

We compare our methods with MUSIC, R-MUSIC, ESPRIT, DOA-AutoEncoder, and SubspaceNet under non-coherent and narrowband circumstances. Switching to the coherent scenario, we add these conventional methods with the SPS technique for

fair comparisons. When the signal is broadband, BB-MUSIC and BB-ESPRIT especially designed for the broadband signal are compared.

4.3. Metrics

For a received signal containing K sources, we define its SNR as,

$$SNR = 10 \log_{10} \left(\frac{\max\{E[|s_1(n)|^2], E[|s_2(n)|^2], \dots, E[|s_K(n)|^2]\}}{E[|v(n)|^2]} \right). \quad (47)$$

The estimation error measure used in this paper is Mean Absolute Error (MAE), which is defined as follows,

$$MAE = \frac{1}{KZ} \sum_{z=1}^Z \sum_{k=1}^K |\theta_{zk} - \hat{\theta}_{zk}|, \quad (48)$$

where θ_{zk} represents the true DOA of the k -th source of the z -th sample, and $\hat{\theta}_{zk}$ represents the predicted DOA of the k -th source of the z -th sample.

4.4. Implementation Details

The base channel number of our Transformer model is empirically set to 128. The input array tokens are first projected into 128 dimensions from M dimensions and proposed L antenna Transformer layers are operated on 128 dimensions. In the feed forward network, the first FC layer upscales the channel number twice to 256 and the last FC layer downscales the channel number to 128. We conducted experiments with varying channel numbers and observed that it had minimal impact on the results. The number of stacked Transformer layers L is set to 3. To optimize the network parameters, we choose the modern automatic differentiation framework PyTorch [56] to implement the proposed Transformer architecture and use the AdamW [57] optimizer with a learning rate of $1e^{-4}$. The conventional methods are implemented in Python and tested several times to ensure numerical stability.

We implement our proposed model on a workstation with an Intel 12-th i9 CPU and two NVIDIA 3090 GPUs. For every group of experiments, the training stage costs around 20 minutes on 10000 simulated samples and the test stage costs only 5 seconds on 2000 simulated samples.

4.5. Main Results

4.5.1. Adaptation of Various SNR

In this section, we evaluate the adaptation of the proposed model against different SNRs. Since the focus is only on the adaptation of the proposed model to varying SNR, we keep parameters except for SNR constant across different experimental groups. The SNR values are sequentially set as -15 dB, -10 dB, -5 dB, 0 dB, and 5 dB. We pay particular attention to the performances of the model under challenging conditions since traditional methods have already demonstrated accurate and effective DOA estimation at high SNR.

The experimental results are shown in Fig. 2. From the observations, we can see that traditional subspace-based methods exhibit excellent performances when SNR is relatively high. This

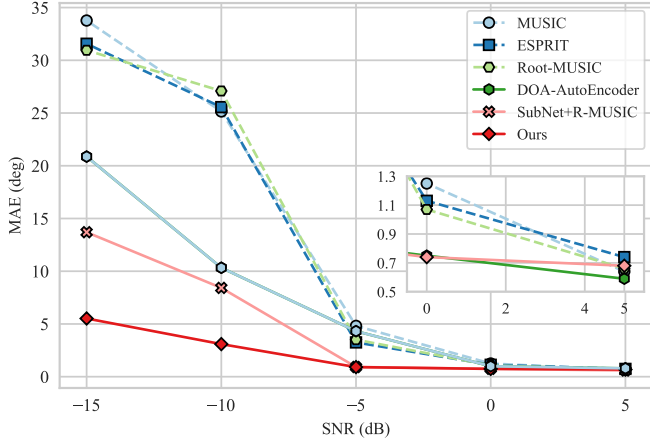


Figure 2: DOA estimation under different SNRs (non-coherent, narrow band, without array errors, $N = 100$, $K = 2$).

is because subspace decomposition methods rely on the processing of the covariance matrix of the received signals. When SNR is high, the boundary between the signal subspace and the noise subspace becomes clear. Subspace methods are able to effectively suppress the influence of the noise subspace, resulting in improved performance. On the other hand, when SNR is low, traditional methods perform worse compared to DL-based methods, since DL-based methods do not rely on specific prior and instead learn complex mappings from raw data, which makes DL-based methods have more extraordinary adaptability to low SNR. It can be observed that our proposed method outperforms the other two DL-based methods under low SNR conditions, demonstrating that our proposed antenna-based attention mechanism is effective in this problem.

4.5.2. Adaptation of Various Number of Sources

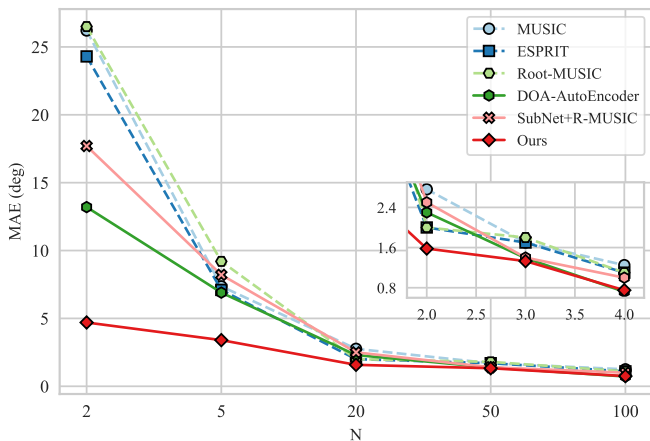


Figure 3: DOA estimation under different numbers of snapshots (SNR = 0 dB, non-coherent, narrow band, without array errors).

In this section, we evaluate the performance of the proposed model under different numbers of sources. We conduct a total of four groups of experiments, where each group keeps the

Table 2: Comparisons with some conventional methods and DL-based methods under different numbers of sources scenario (SNR = 0 dB, $N = 100$, non-coherent, narrow band, without array errors) Performances are reported by absolute error (deg).

methods	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
MUSIC	0.67	1.25	2.66	4.98	6.90
ESPRIT	0.58	1.13	2.77	4.71	5.16
Root-MUSIC	0.57	1.07	2.41	3.19	4.90
DOA-AutoEncoder	0.62	0.97	2.09	2.76	4.05
SubNet+R-MUSIC	0.52	0.84	1.67	2.13	3.80
Ours	0.46	0.75	1.51	1.80	3.22

parameters as same as the basic setting, except for the number of sources. The number of sources, denoted as K , is sequentially set as 1, 2, 3, 4, and 5. In practical scenarios, super-resolution DOA estimation typically requires $M > K$. In our experiments, we set the number of array elements, M , as 10. Theoretically, the possible number of sources ranges from 1 to 9, but for simplicity and focus on the main cases, we set the number of sources from 1 to 5.

The experimental results are shown in Tab. 2. It can be observed that traditional methods often yield lower accuracy in handling multi-source problems under the poor condition of 0 dB. In contrast, the DOA-Autoencoder and Subspacenet methods outperform traditional methods, which can be attributed to the incorporation of deep modules that enable the model to learn multiple features globally rather than relying heavily on a single feature. Furthermore, the experimental results indicate that our proposed model exhibits the best adaptability to multi-source problems. The average absolute errors in the cases of 1, 2, 3, and 4 sources are 0.46, 0.75, 1.51, and 1.80 degrees, respectively, which again highlights the strong adaptability of our model to multi-source problems. In order to visually examine the performance of our model, we conducted angle scanning tests for two scenarios: $K = 1$ and $K = 3$. Fig. 4(a) and Fig. 4(b) represent the prediction results and errors of our model for the $K = 1$ scenario, where the angle scanning range for a single source is set from -60 to 60 degrees. Fig. 4(c) and Fig. 4(d) represent the prediction results and errors of our model for the $K = 3$ scenario, where we set a constant angular separation of 10 degrees between each source. It can be observed that the overall errors roughly align with the results presented in Tab. 2, which also validates the effectiveness of the proposed antenna-based attention mechanism in addressing this problem.

4.5.3. Adaptation of Various Snapshots

In this section, we explore the impact of different snapshot numbers N on the DOA estimation performance. To assess the adaptability of the proposed model to snapshot numbers, we conduct four groups of experiments where parameters, except for the snapshot number, remained constant. The snapshot numbers are sequentially set as 100, 50, 20, 5, and 2.

The experimental results are shown in Fig. 3. When the number of snapshots is high, traditional methods exhibit excellent performance, which is due to the increased time-domain reso-

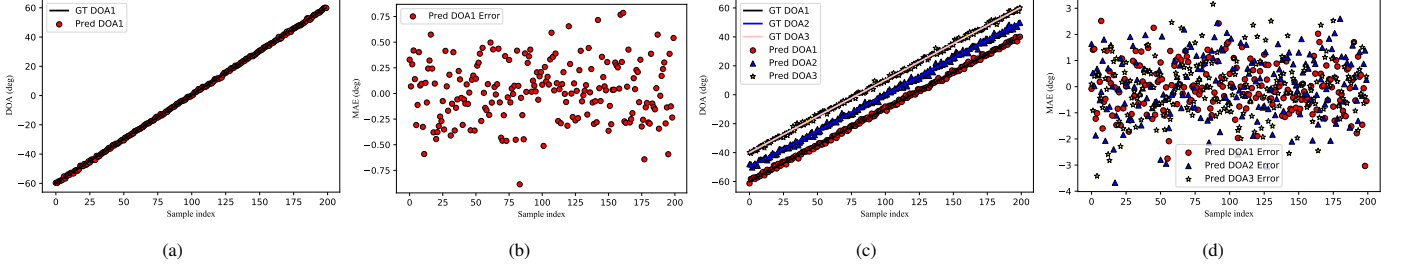


Figure 4: DOA estimation with well trained deep Transformer model under $K = 1$ and 3 scenarios. (a) is the predicted DOA and the groundtruth DOA under the $K = 1$ setting, (b) is the absolute errors under the $K = 1$ setting. (c) is the predicted DOA and the groundtruth DOA under the $K = 3$ setting, (d) is the absolute errors under the $K = 3$ setting. “GT DOA” denotes the groundtruth DOA and “Pred DOA” represents the predicted DOA estimated by the network.

lution provided by a high number of snapshots. However, when the number of snapshots is low, DL-based methods demonstrate good adaptability. As the snapshot number decreases from 10 to 2, the performance of traditional methods sharply declines. This is due to the inability of temporal averaging of the covariance matrix to accurately replace sample averaging in low snapshot scenarios. The Subspacenet method, which relies on recovering the covariance matrix, also performs poorly in low snapshot numbers because of this limitation. Our method demonstrates the best performance in low snapshot conditions, which further confirms the effectiveness of the antenna-based attention mechanism.

4.5.4. Adaptation of Coherency

In this section, we explore the impact of signal coherency on the DOA estimation performance. Traditional subspace methods experience a decline in estimation performance when there is coherence among the sources, as the rank of the covariance matrix decreases, making spatial decomposition infeasible. In order to demonstrate the ability of the proposed model to handle coherent signals, we conduct a total of four major groups, each consisting of 16 subgroups of experiments. In each major group, the number of snapshots is set to 50, 20, 5, and 2, respectively, and the SNR values are sequentially set to -5 dB, -3 dB, 0 dB, and 5 dB, while the remaining parameters are kept the same as the basic setting.

The coherent source modeling method we adopt is as follows: taking two signals $s_i(n)$ and $s_k(n)$ as an example, their correlation coefficient is defined as,

$$\rho_{ik} = \frac{E[s_i(n)s_k^*(n)]}{\sqrt{E[|s_i(n)|^2]E[|s_k(n)|^2]}}. \quad (49)$$

From the Schwarz inequality, we know that,

$$|\rho_{ik}| \leq 1. \quad (50)$$

Therefore, the correlation between signals is defined as follows,

$$\begin{cases} |\rho_{ik}| = 0, & s_i(t), s_k(t) \text{ irrelevant} \\ 0 < |\rho_{ik}| < 1, & s_i(t), s_k(t) \text{ relevant} \\ |\rho_{ik}| = 1, & s_i(t), s_k(t) \text{ coherent} \end{cases} \quad (51)$$

This paper mainly considers the situation when the signal sources are coherent ($|\rho_{ik}| = 1$). Specifically, when there are

K signal sources that are coherent, the array reception model is,

$$\mathbf{X} = \mathbf{A}\boldsymbol{\rho}s_1(n) + \mathbf{V}, \quad (52)$$

where,

$$\boldsymbol{\rho} = [|\rho_1|e^{j\Delta\phi_1}, |\rho_2|e^{j\Delta\phi_2}, \dots, |\rho_k|e^{j\Delta\phi_k}, \dots, |\rho_K|e^{j\Delta\phi_K}] \in \mathbb{C}^{K \times 1}, \quad (53)$$

$$|\rho_k| = 1, k = 1, 2, \dots, K, \quad (54)$$

$$|\phi_k| \in [-30^\circ, 30^\circ], k = 1, 2, \dots, K. \quad (55)$$

The experimental results are shown in Fig. 5. It can be observed that the SPS-MUSIC, SPS-ESPRIT, and SPS-R-MUSIC methods perform on par with the DL-based methods in terms of performance at high snapshot numbers and high SNR, even surpassing the DOA-Autoencoder method in some cases. This is because under relatively ideal conditions, traditional methods can be considered to provide analytical solutions, and thus their performance exceeds prediction-based DL methods. Moreover, our proposed model demonstrates superior performance at low SNR or low snapshot numbers, which is caused by the method we use does not entirely rely on the construction of the covariance matrix itself, but learns useful features about DOA from the original data. In a word, the overall experimental results show that the antenna-based attention mechanism proposed in our study possesses excellent processing capabilities for coherent signals.

4.5.5. Adaptation of Array errors

In this section, we investigate the impact of array errors on the DOA estimation performance. The array errors considered include mutual coupling errors, channel gain/phase inconsistencies errors, element position errors, and directional diagram errors. Accurately modeling these errors can be challenging, therefore, we adopt a simplified representation of the errors based on the approach proposed in [42]. The mutual coupling errors vector is represented as follows,

$$\mathbf{c}_{err} = \alpha \times [c_1, c_2, \dots, c_M]^T, \quad (56)$$

where c_i is a complex number with amplitude ranging from 0.8 to 1 and phase ranging from -30 to 30 degrees. The specific values can be chosen during the experiments. α is the errors strength coefficient, which controls the severity of the errors.

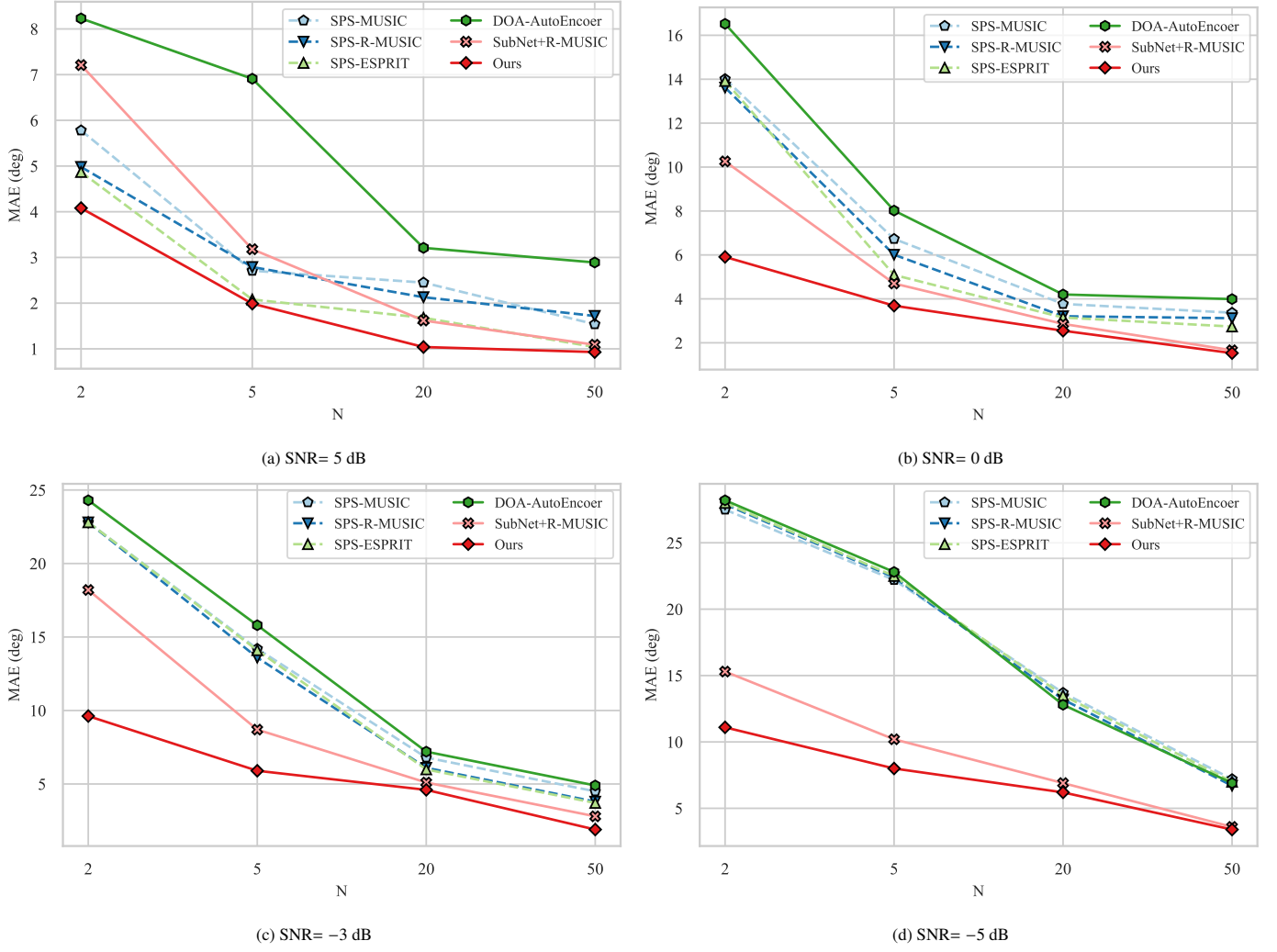


Figure 5: Comparisons with other methods under the coherent-source scenario. Conventional methods are presented in dash lines, while DL-based methods are plotted in solid lines.

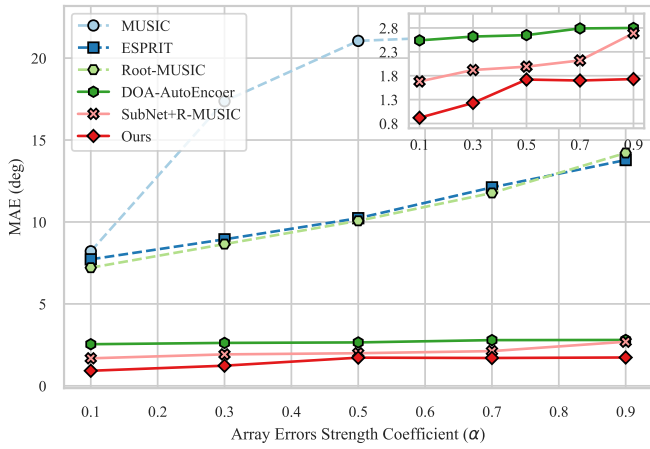


Figure 6: DOA estimation under different array errors strength coefficients α (SNR = 0 dB, non-coherent, narrow band, and $N = 100$).

Similarly, the gain/phase inconsistencies errors matrix is formulated as follows,

$$\mathbf{L}_{err} = \alpha \times \text{diag}[l_1, l_2, \dots, l_M], \quad (57)$$

where l_i is a complex number with amplitude ranging from 0.8 to 1 and phase ranging from -30 to 30 degrees. The specific values can be chosen during the experiments. The element position errors vector is formed as follows,

$$\mathbf{b}_{err} = \alpha \times [b_1, b_2, \dots, b_M]^T \times d, \quad (58)$$

where $b_i \sim \mathcal{U}(-\alpha, \alpha)$. The directional diagram error vector can be denoted as,

$$\mathbf{p}_{err} = \alpha \times [p_1, p_2, \dots, p_M]^T, \quad (59)$$

where p_i is a complex number with amplitude ranging from 0.8 to 1 and phase ranging from -30 to 30 degrees. The specific values can be chosen during the experiments. We believe these simplified approaches are reasonable because the deep model does not rely on specific prior information about array errors.

Then, we sequentially set α from 0.1 to 0.9 to explore the adaptability of the model to array errors. The experimental results are shown in Fig. 6. It can be observed that as α increases, the estimation errors of traditional methods exhibit a noticeable upward trend. On the other hand, the errors of the DL-based methods do not show significant changes with increasing α . This is because the deep model does not rely on specific array prior assumptions, in other words, DL-based methods do not require prior knowledge of the type and extent of errors when performing array error correction, making it robust to various array errors. Our proposed model achieves the best performance, which further confirms the effectiveness of the antenna-based attention mechanism.

4.5.6. Adaptation of Broadband

In this section, we consider the performance of the model on broadband signals. The modeling of broadband signals is based on the approach proposed in [43], which can be represented as follows,

$$s_k(t) = \frac{1}{L} \sum_{l=0}^{L-1} s_{k,l} e^{2\pi j l \frac{B_f}{L f_s} t}, \quad (60)$$

where L represents the number of subcarriers, $s_{k,l}$ represents the l^{th} subcarrier of the k^{th} signal, which is independently modulated with a zero-mean complex Gaussian distribution with unit variance. f_s represents the sampling frequency and B_f means the signal bandwidth. In the experiments, we set $L = 500$, $f_s = 500\text{Hz}$, $B_f = 500\text{Hz}$, and the snapshot numbers are sequentially set to 500, 200, 50, and 20. The remaining parameters were kept the same as the basic setting.

The experimental results are shown in Fig. 7. We compare the BB-MUSIC method, and the BB-ESPRIT method for broadband sources, as well as the three DL-based methods. It can be seen that the performance of the BB-MUSIC and BB-ESPRIT is relatively good at high snapshot numbers, even surpassing the DOA-Autoencoder method in some cases. However, as the snapshot numbers decrease, the limitations of traditional methods become apparent. This is because BB-MUSIC and BB-ESPRIT require decomposing the broadband signal into several narrowband signals. When the snapshot number is low, the sampled data does not contain sufficient frequency domain information, resulting in information loss and a decrease in DOA performance. In contrast, DL-based models significantly exhibit high performance on broadband signal estimation, particularly in low snapshot environments. We can see that our proposed model performs best in this problem which stems from the careful design of our model structure for this specific problem, including the antenna-based attention mechanism for joint feature extraction.

4.6. Ablation Study

The ablation study in this section explores the Antenna-based Attention and High Dimensional Bound of Attention respectively, aiming to prove that our default settings of the proposed model are reasonable and effective.

4.6.1. Antenna-based Attention

We conduct the ablation study on our proposed antenna-based attention by implementing three variant models on the baseline. All variant models are trained to the convergence. The three variant models are listed as follows,

1. Only traditional spatial attention (i.e., only \mathcal{A}_S),
2. Only real-part attention (i.e., only \mathcal{A}_R),
3. Only imaginary-part attention (i.e., only \mathcal{A}_I),
4. Our proposed model.

Table 3: Ablation study on our proposed antenna-based attention. ✓ denotes that the corresponding attention is employed. The data in the table represents the absolute error.

\mathcal{A}_S	\mathcal{A}_R	\mathcal{A}_I	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
✓			1.26	1.59	2.43	2.88	5.46
	✓		0.82	1.21	1.93	2.29	3.80
		✓	0.55	0.88	1.82	2.01	3.52
	✓	✓	0.46	0.75	1.51	1.80	3.22

The ablation results are reported in Tab. 3.3. It's clear to see that using traditional spatial attention would harm the estimation and only using \mathcal{A}_R or using \mathcal{A}_I would cause performance degradation. Using spatial attention \mathcal{A}_S can not model the covariance matrix \mathcal{R} and benefit from its task-related property. Moreover, only performing one of the real-part \mathcal{A}_R or imaginary-part \mathcal{A}_I attention cannot estimate DOA well. By using the proposed antenna-based attention, the model can achieve state-of-the-art (SOTA) performance.

4.6.2. High Dimensional Bound of Attention

In Sec. 3.4, we discuss the attention map (i.e., \mathcal{A}_R and \mathcal{A}_I) should be bounded in high dimension to ensure the lossless information of the signal covariance matrix. We choose to add an auxiliary reconstruction loss to restrict the attention map in high-dimensional feature space. To verify the effectiveness of the restriction, we design the ablation model without using the reconstruction loss. After training it on the basic setting, we find that the performance without reconstruction loss is degraded. The absolute error rises from 0.75 to 0.89, which shows that the high-dimensional bound is necessary.

4.7. Discussion

This section mainly discusses the performance of our proposed model in various more complex scenarios that may be encountered, including Complex Situation, Real World Signal, Generalization Ability for Unseen Scenario and Blind Number of Sources, aiming to verify the adaptability and robustness of our model.

4.7.1. Complex Situation

This section explores the adaptability of models to various complex situations. The signals in real-world scenarios can be complex and influenced by various factors. Therefore, it is necessary to simultaneously consider the various situations discussed in Sec. 4.

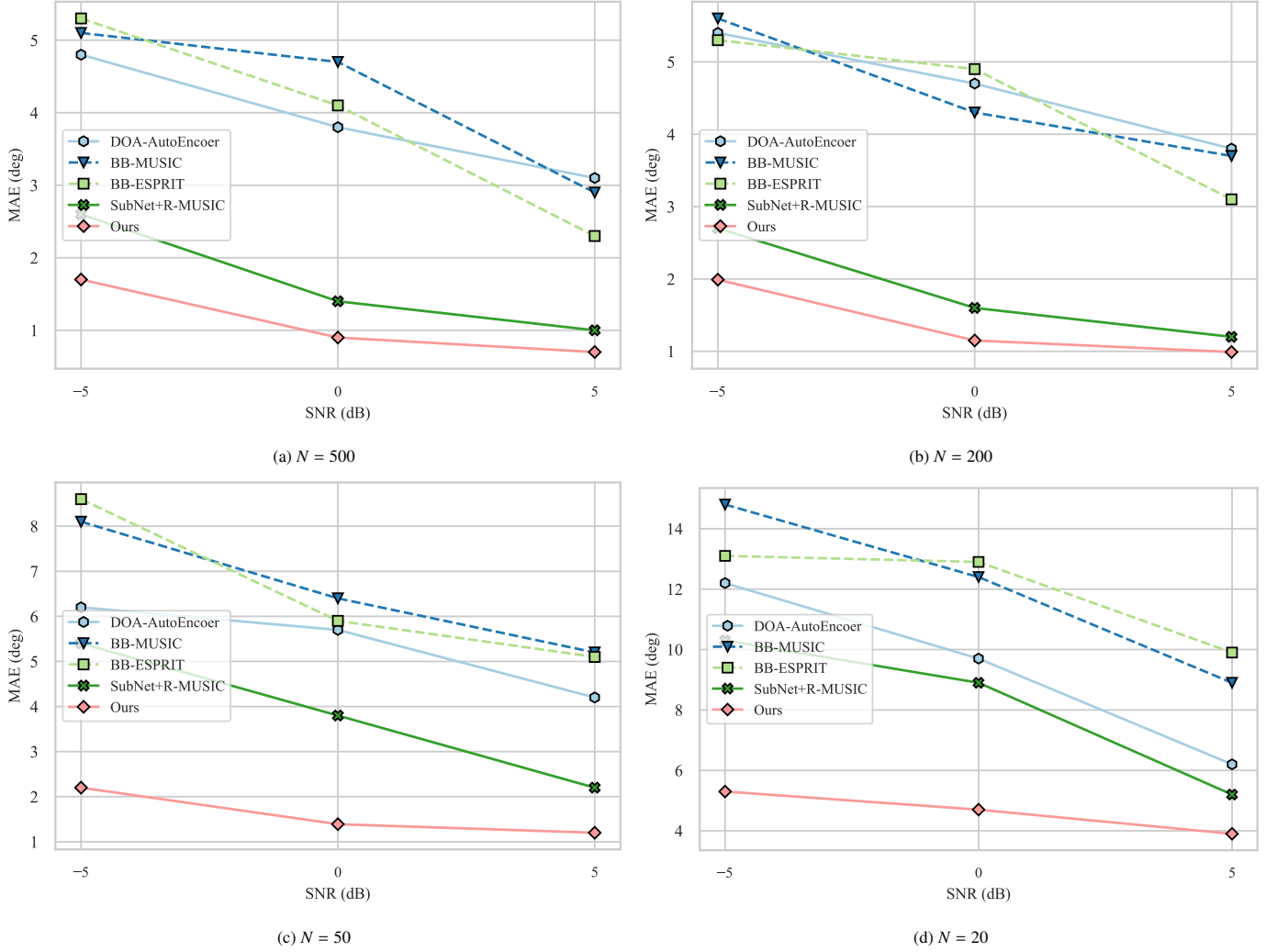


Figure 7: Comparisons with other methods under the broadband-source scenario. Conventional methods are presented in dash lines, while DL-based methods are plotted in solid lines.

In this section, we conducted a total of six groups of experiments to explore the adaptability of the model to complex situations. In these groups, we gradually make the experimental conditions more challenging while keeping the number of snapshots constant, because the number of snapshots is often controllable even in complex situations. The experimental setup and results are shown in Tab. 4. From the comparison of the first and second rows, we can see that the introduction of broadband signals leads to the degradation of model performance. This is because broadband signals contain more complex frequency domain information, which may pose learning difficulties for the model with a limited number of snapshots. From the comparison of the second and third rows, we can see that increasing array errors leads to degradation in model performance, but the impact is not significant. This is consistent with the analysis in Sec 4, as deep models do not rely on specific prior information about the array errors during the learning process. From the comparison of the third and fourth rows, we can observe that decreasing the SNR results in a decrease in model performance. We conducted a dedicated analysis on this aspect in

Sec 4. From the comparison of the fourth and fifth rows, it can be seen that increasing the number of signal sources leads to a rapid decline in model performance. Going from $K = 2$ to $K = 3$ nearly doubles the absolute estimation error of the model. The sixth row represents a set of extreme conditions that we set as a control. Under these conditions, the model's estimation error reaches 5.56 degrees, which shows that our model has great adaptability to complex scenarios.

4.7.2. Real World Signal

In this section, we validate the proposed model on real-world data, which consists of single-frequency signals with a frequency of 30 KHz, snapshots $N = 512$, SNR = 5 dB, number of array elements $M = 5$, and a uniform circular array. Since the array structure of the real-world data is a uniform circular array, we first generate simulated data with a uniform circular array configuration for training the model. Subsequently, we perform fine tune by using 15 samples from the real-world data as the training set, while the remaining 5 samples are used for testing. Specifically, due to the small amount of measured

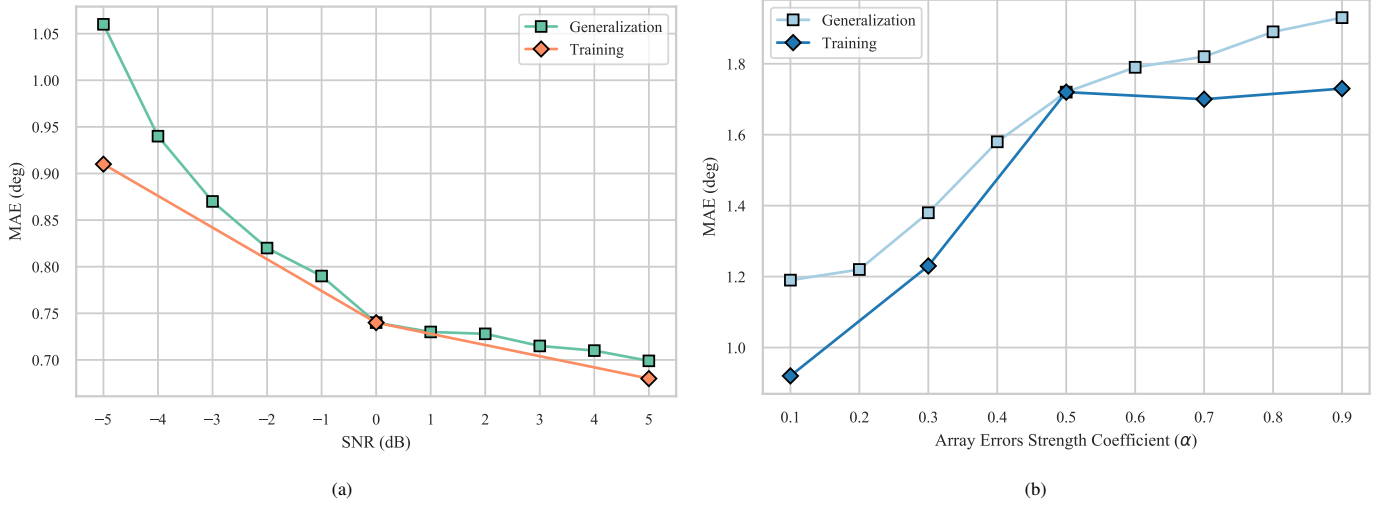


Figure 8: Generalization ability of our proposed method. (a) denotes that model is trained on SNR= 0 dB and tested on different SNRs. (b) represents that the model is trained on $\alpha = 0.5$ and tested on different α .

Table 4: Performances of our proposed Transformer under extremely complex scenarios.

SNR	Snapshots	Array errors	Coherent	Broadband	Number sources	Abs. Error (deg)
5 dB	$N = 500$	$\alpha = 0.1$	✓		$K = 2$	0.32
5 dB	$N = 500$	$\alpha = 0.1$	✓	✓	$K = 2$	1.03
5 dB	$N = 500$	$\alpha = 0.3$	✓	✓	$K = 2$	1.16
0 dB	$N = 500$	$\alpha = 0.3$	✓	✓	$K = 2$	1.49
0 dB	$N = 500$	$\alpha = 0.3$	✓	✓	$K = 3$	2.82
0 dB	$N = 500$	$\alpha = 0.5$	✓	✓	$K = 5$	5.56

Table 5: Comparisons between our method and a widely-used conventional algorithm MUSIC on the real-world dataset which obtains samples with SNR= 0 dB.

methods	DOA degree				
	-60	-30	0	30	60
MUSIC	2.00	1.20	3.50	0.67	2.20
Ours	1.37	0.83	1.52	0.43	1.31

data, updating all parameters of the model can easily lead to overfitting. Therefore, we freeze all parameters except the five matrices \mathbf{W}_R^Q , \mathbf{W}_R^K , \mathbf{W}_I^Q , \mathbf{W}_I^K and \mathbf{W}^V [58], and use the measured data to fine-tune the Transformer model that has been trained on the simulated data, so that the model can adapt to the measured data. The results are shown in Tab. 5. It can be observed that our method demonstrates more stable performance on real-world data compared to MUSIC. This indicates that our method exhibits stronger adaptability to variations and noise in the real world, enabling more accurate DOA estimation. Traditional methods may be more susceptible to disturbances from non-ideal factors when dealing with real-world data, leading to less stable and accurate estimation results. However, our proposed method overcomes these challenges, enhancing the robustness and accuracy of DOA estimation.

4.7.3. Generalization Ability for Unseen Scenario

In this section, we conduct generalization tests on the proposed method to explore how the model can extend its learning outcomes to scenarios not included in the training dataset. This experiment is necessary because, in practical applications, it is often impossible to precisely model complex scenarios, which inevitably leads to the operation of the model in new environments. Two groups of experiments are conducted to investigate the model's generalization ability concerning SNR and array errors, respectively. For the former, we trained the model only with an SNR of 0 dB and performed generalization tests on SNR ranging from -5 dB to 5 dB. For the latter, we trained the model with $\alpha = 0.5$ and conducted generalization tests with α ranging from 0.1 to 0.9.

The experimental results are shown in Fig. 8. From Fig. 8(a), we can see that the performance degradation rate at which the absolute error decreases with increasing SNR is smaller than the rate at which the absolute error increases with decreasing SNR. This is because, at lower SNR, the model not only experiences difficulties in learning due to the low SNR itself, but generalization also introduces more errors. Similar conclusions can be drawn from Fig. 8(b), where the performance degradation rate at which the absolute error increases with increasing array errors intensity is greater than the rate at which the absolute error decreases with decreasing array errors intensity. These two groups of experiments demonstrate that our model has a certain

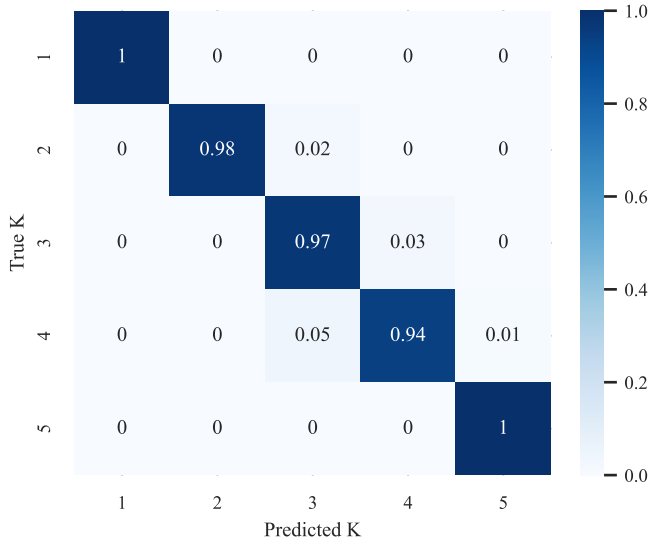


Figure 9: Confusion matrix of proposed classification network.(SNR= 0 dB, $N = 100$, non-coherent, narrow band, and without array errors)

level of generalization ability and can handle variations in scenarios. Although its direct generalization performance is not as good as the performance achieved when tested under the same conditions as training, overall, its performance is still satisfactory.

4.7.4. Blind Number of Sources

In practical DOA estimation, the number of signal sources may not be known, while our model requires information about the number of sources for application. Therefore, in this section, we propose a framework for estimating the number of signal sources to assist our DOA estimation model. We utilize the same model architecture as the DOA estimation framework for handling the source number estimation problem, with the only difference being the replacement of the regression head in the model's backend with a classification head and a reduction in the number of Transformer layers.

The experimental setup is the same as the basic setting, and the resulting confusion matrix is shown in Fig. 9. Each element m_{ij} in the confusion matrix represents the probability of predicting the number of true signal sources as i when the actual number of sources is j . The element to the right of the diagonal represents the false positive (FP) rate, and the value to the left of the diagonal represents the false negative (FN) rate. Under the conditions of SNR 0 dB and 100 snapshots, such results are satisfactory because the problem that estimating the number of multiple sources in adverse environments is often challenging to solve. After estimating the number of sources, we can employ the corresponding trained regression model to estimate the DOA in a cascade manner.

5. CONCLUSION

In this paper, we propose a novel Transformer model to solve the DOA estimation problem. Our method enhances the origi-

nal Transformer method by introducing an antenna-based attention mechanism specially designed for DOA estimation. These changes allow the model to be guided to learn to a certain extent and pay more attention to the information that is really useful for DOA estimation, which is proven to be effective in DOA estimation tasks. The working principles of these changes are mathematically derivable, thus providing a degree of interpretability to our model. Compared to traditional methods and other DL-based approaches, our proposed method demonstrates superior performance. It is capable of handling scenarios with a low SNR, a limited number of snapshots, multiple signal sources, coherent sources, broadband sources, and array errors. It also exhibits excellent adaptability to extremely complex scenarios and unknown scenarios. These abilities show that our method has strong practicality compared with traditional methods and good interpretability compared with previous DL-based methods.

6. Acknowledge

This paper is supported by National Natural Science Foundation of China (NSFC) No.62031007 and No.62231006.

References

- [1] Ming Jin, Guisheng Liao, and Jun Li. Joint dod and doa estimation for bistatic mimo radar. *Signal processing*, 89(2):244–251, 2009. 1
- [2] Ali Gorcin and Huseyin Arslan. A two-antenna single rf front-end doa estimation system for wireless communications signals. *IEEE transactions on antennas and propagation*, 62(10):5321–5333, 2014. 1
- [3] Xinping Lin, Xiaofei Zhang, Lang He, and Wang Zheng. Multiple emitters localization by uav with nested linear array: System scheme and 2d-doa estimation algorithm. *China Communications*, 17(3):117–130, 2020. 1
- [4] Jianguo Huang, Lijie Zhang, Qunfei Zhang, Yong Jin, and Min Jiang. Performance analysis of doa estimation for mimo sonar based on experiments. In *IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 269–272. IEEE, 2009. 1
- [5] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996. 1
- [6] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE Assp Magazine*, 5(2):4–24, 1988. 1
- [7] Sergiy A Vorobyov, Alex B Gershman, and Zhi-Quan Luo. Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem. *IEEE Transactions on Signal Processing*, 51(2):313–324, 2003. 1
- [8] Petre Stoica and Alex B Gershman. Maximum-likelihood doa estimation by data-supported grid search. *IEEE Signal Processing Letters*, 6(10):273–275, 1999. 1
- [9] Bo Tang, Jun Tang, Yu Zhang, and Zhidong Zheng. Maximum likelihood estimation of dod and doa for bistatic mimo radar. *Signal Processing*, 93(5):1349–1357, 2013. 1
- [10] Michael I Miller and Daniel R Fuhrmann. Maximum-likelihood narrow-band direction finding and the em algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1560–1577, 1990. 1
- [11] Petre Stoica and Arye Nehorai. Music, maximum likelihood, and cramer-rao bound. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(5):720–741, 1989. 1
- [12] Richard Roy and Thomas Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989. 1, 4, 2
- [13] A-J Van Der Veen, ED F Deprettere, and A Lee Swindlehurst. Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308, 1993. 1

- [14] Tsung-Hsien Liu and Jerry M Mendel. A subspace-based direction finding algorithm using fractional lower order statistics. *IEEE Transactions on Signal Processing*, 49(8):1605–1613, 2001. 1
- [15] Feng-Gang Yan, Liu Shuai, Jun Wang, Jun Shi, and Ming Jin. Real-valued root-music for doa estimation with reduced-dimension evd/svd computation. *Signal Processing*, 152:1–12, 2018. 1
- [16] Zhang-Meng Liu, Zhi-Tao Huang, and Yi-Yu Zhou. Sparsity-inducing direction finding for narrowband and wideband signals based on array covariance vectors. *IEEE Transactions on Wireless Communications*, 12(8):1–12, 2013. 1
- [17] Yuexian Wang, Xin Yang, Jian Xie, Ling Wang, and Brian W-H Ng. Sparsity-inducing doa estimation of coherent signals under the coexistence of mutual coupling and nonuniform noise. *IEEE Access*, 7:40271–40278, 2019. 1
- [18] Alice Delmer, Anne Ferréol, and Pascal Larzabal. L0 regularization parameter for sparse doa estimation of coherent signals with modeling errors. *Signal Processing*, 209:109006, 2023. 1
- [19] Jin Wang, Yongjun Zhao, and Zhigang Wang. A music like doa estimation method for signals with low snr. In *Global Symposium on Millimeter Waves (BSMM)*, pages 321–324. IEEE, 2008. 1
- [20] Elias Aboutanios, Aboulnasr Hassanien, Moeness G Amin, and Abdelhak M Zoubir. Fast iterative interpolated beamforming for accurate single-snapshot doa estimation. *IEEE Geoscience and Remote Sensing Letters*, 14(4):574–578, 2017. 1
- [21] Hongyi Wang, KJ Ray Liu, and Henry Anderson. Spatial smoothing for arrays with arbitrary geometry. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–509. IEEE, 1994. 1
- [22] Qing Shen, Wei Liu, Wei Cui, and Siliang Wu. Underdetermined doa estimation under the compressive sensing framework: A review. *IEEE Access*, 4:8865–8878, 2016. 1
- [23] Yeo-Sun Yoon, Lance M Kaplan, and James H McClellan. Tops: New doa estimator for wideband signals. *IEEE Transactions on Signal processing*, 54(6):1977–1989, 2006. 1, 4.2
- [24] Benjamin Friedlander and Anthony J Weiss. Direction finding in the presence of mutual coupling. *IEEE Transactions on Antennas and Propagation*, 39(3):273–284, 1991. 1
- [25] Jisheng Dai, Dean Zhao, and Xiaofu Ji. A sparse representation method for doa estimation with unknown mutual coupling. *IEEE Antennas and Wireless Propagation Letters*, 11:1210–1213, 2012. 1
- [26] Massimo Donelli, Federico Viani, Paolo Rocca, and Andrea Massa. An innovative multiresolution approach for doa estimation based on a support vector classification. *IEEE Transactions on Antennas and Propagation*, 57(8):2279–2292, 2009. 1
- [27] Michał Tarkowski and L Kulas. Rss-based doa estimation for espar antennas using support vector machine. *IEEE Antennas and Wireless Propagation Letters*, 18(4):561–565, 2019. 1
- [28] Andrea Randazzo, Mohamed A Abou-Khousa, Matteo Pastorino, and R Zoughi. Direction of arrival estimation based on support vector regression: Experimental validation and comparison with music. *IEEE Antennas and Wireless propagation letters*, 6:379–382, 2007. 1
- [29] Matteo Pastorino and Andrea Randazzo. A smart antenna system for direction of arrival estimation based on a support vector regression. *IEEE Transactions on Antennas and Propagation*, 53(7):2161–2168, 2005. 1
- [30] Ahmed H El Zooghby, Christos G Christodoulou, and Michael Georgiopoulos. Performance of radial-basis function networks for direction of arrival estimation with antenna arrays. *IEEE Transactions on Antennas and Propagation*, 45(11):1611–1617, 1997. 1
- [31] Saber Helmy Zainud-Deen, HA Malhat, Kamal Hassan Awadalla, and ES El-Hadad. Direction of arrival and state of polarization estimation using radial basis function neural network (rbfnn). In *National Radio Science Conference (NRSC)*, pages 1–8. IEEE, 2008. 1
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1
- [33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 1
- [34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 1
- [35] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 1
- [36] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 1
- [37] He Cao, Jianan Wang, Tianhe Ren, Xianbiao Qi, Yihao Chen, Yuan Yao, and Lei Zhang. Exploring vision transformers as diffusion learners. *arXiv preprint arXiv:2212.13771*, 2022. 1
- [38] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1
- [39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1
- [40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021. 1
- [41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1
- [42] Zhang-Meng Liu, Chenwei Zhang, and S Yu Philip. Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections. *IEEE Transactions on Antennas and Propagation*, 66(12):7315–7327, 2018. 1, 3.2, 4.2, 4.5.5
- [43] Dor H Shmuel, Julian P Merkofer, Guy Revach, Ruud JG van Sloun, and Nir Shlezinger. Subspacenet: Deep learning-aided subspace methods for doa estimation. *arXiv preprint arXiv:2306.02271*, 2023. 1, 3.2, 4.2, 4.5.6
- [44] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advance on Neural Information Processing Systems (NeurIPS)*, 2017. 1, 3.2
- [45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1
- [46] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan Loddon Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *ArXiv*, abs/2102.04306, 2021. 1
- [47] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision (ECCV) Workshops*, 2021. 1
- [48] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, P. Abbeel, A. Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advance on Neural Information Processing Systems (NeurIPS)*, 2021. 1
- [49] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelu). *arXiv: Learning*, 2016. 3.2
- [50] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 3.4
- [51] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015. 3.4
- [52] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986. 4.2
- [53] Arthur Barabell. Improving the resolution performance of eigenstructure-based direction-finding algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 8, pages 336–339. IEEE, 1983. 4.2
- [54] S Unnikrishna Pillai and Byung Ho Kwon. Forward/backward spatial smoothing techniques for coherent signal identification. *IEEE Transac-*

- tions on Acoustics, Speech, and Signal Processing, 37(1):8–15, 1989. 4.2
- [55] Muhammad Faisal Khan and Muhammad Tufail. Performance analysis of esprit algorithm on a single broadband signal. In *2009 International Conference on Emerging Technologies*, pages 310–314. IEEE, 2009. 4.2
 - [56] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 4.4
 - [57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2017. 4.4
 - [58] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *European Conference on Computer Vision (ECCV)*, pages 497–515. Springer, 2022. 4.7.2