

# An Off-Grid DOA Estimation Method Based on a Frequency-Domain ViT

He Zheng, Guimei Zheng, *Member, IEEE*, Fangqing Wen, *Senior Member, IEEE*, Yuwei Song and Feilong Lv

**Abstract**—In this letter, we propose a deep learning-based off-grid Direction of Arrival (DOA) estimation method for low Signal-to-Noise Ratio (SNR) scenarios. Specifically, we develop a dual-branch neural network with residual connections that processes frequency-domain features, consisting of a coarse classification branch and a fine regression branch. The classification branch employs a multi-label approach to obtain on-grid results, while the regression branch predicts the residual between the classification outputs and ground-truth angles. This structural design effectively leverages classification results to avoid convergence difficulties associated with direct off-grid angle prediction, thereby enhancing DOA estimation accuracy. Simulation results demonstrate that under low SNR conditions, the proposed method outperforms existing approaches, including both classical model-based and other deep learning-based methods.

**Index Terms**—DOA estimation, Vision Transformer (ViT), low signal-to-noise ratio (SNR), Deep Learning.

## I. INTRODUCTION

Direction of Arrival (DOA) estimation is a crucial component of array signal processing, with extensive applications in fields such as communications, radar, and the Internet of Things. In recent years, with the continuous deepening of research, numerous methodologies have been proposed, including Multiple Signal Classification (MUSIC) [1], Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [2], and compressed sensing-based approaches [3]. The key to these methods lies in deriving mathematical relationships and solution formulas between signal parameters and received data through mathematical models of array-received signals [4], [5]. By relying on these mathematical models to obtain estimation values, they are also classified as model-driven approaches. Although model-driven

This work was supported by National Natural Science Foundation of China under Grant 62401619, 62271286, and the Open Fund of Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering (2025SDSJ05). (Corresponding author: Fangqing Wen and Guimei Zheng).

Fangqing Wen is with the College of Computer and Information Technology and the Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, Hubei, China (e-mail: wenfangqing@ctgu.edu.cn). Also, he is with the Faculty of Data Science, City University of Macau, Macau, China.

He Zheng, Feilong Lv, Yuwei Song and Guimei Zheng are with Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China

methods such as MUSIC [1], ESPRIT [2], and compressive sensing-based approaches [3] exhibit excellent estimation performance under ideal conditions, their effectiveness heavily relies on the accuracy of array models and signal assumptions. Under non-ideal conditions such as low signal-to-noise ratio (SNR), limited snapshots, or coherent sources, these methods often suffer significant performance degradation due to model mismatch [21–23]. For example, Johnson et al. [21] provides a theoretical analysis of the resolution limit of MUSIC under low SNR; Li et al. [22] highlights the high sensitivity of ESPRIT to array errors; and Wax et al. further demonstrates that traditional methods require additional decorrelation processing in coherent source scenarios, which increases computational complexity and uncertainty [23]. Additionally, Dai et al. points out that the computational complexity of sparse Bayesian methods is excessively high, making them unsuitable for real-time processing [20].

In recent years, deep learning has profoundly impacted computer vision and Natural Language Processing (NLP) [7], prompting research into its array signal processing applications like DOA estimation [6], [8], [9], [10], [11], [17]. Unlike model-driven methods that rely on idealized mathematical assumptions—which degrade in low-SNR or correlated source scenarios, data-driven approaches optimize neural networks through large datasets to extract features and model nonlinear mappings. These methods benefit from the Scaling Law [12]: performance improves with increased model parameters and training data, enhancing generalization across diverse scenarios. While training is computationally intensive, optimized architectures enable efficient inference. This paradigm shift offers robustness in complex environments where traditional parametric models fail, addressing limitations through data scalability and adaptive feature learning rather than rigid mathematical formulations. Papageorgiou et al. proposes a Convolutional Neural Network (CNN) based DOA estimation method that extracts covariance matrix features through a series of convolutional layers, ultimately generating angle-wise probability spectra for DOA realization [13]. Guo et al. enhances estimation accuracy under low SNR by first using a convolutional network for feature extraction, then integrating a classification head into a Vision Transformer (ViT) [14], [15] variant called DCT-ViT to refine categorization precision [16].

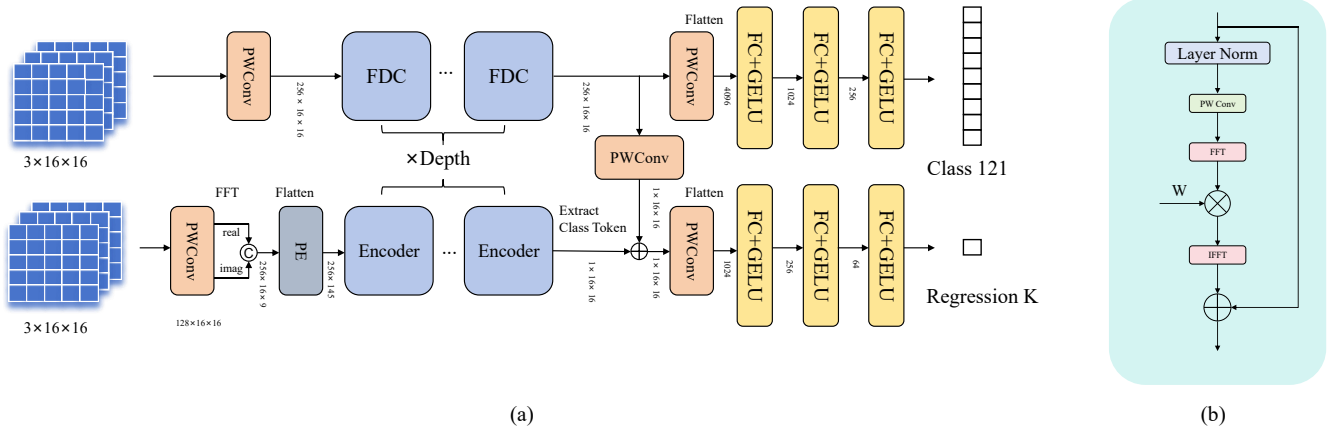


Fig. 1 FViT Network Architecture Diagram. (a) Structure of FViT. (b) Structure of FDC

To address the inherent grid discretization errors in multi-label classification-based DOA estimation methods, this paper proposes a novel off-grid estimation framework. By integrating classification and regression branches through residual connections, the framework first performs coarse angle classification to narrow the direction range, then employs a regression module to predict the continuous residual between the coarse estimate and the true angle. This decoupled design effectively eliminates quantization errors while preserving the robustness of data-driven feature extraction, yielding strong performance in challenging scenarios such as low SNR, limited snapshots, and correlated sources. Integrating FFT into the neural network embeds the physical principles of traditional beamforming—particularly the Fourier relationship between spatial signal sampling and its wavenumber spectrum—as a learnable inductive bias. This guides feature learning into the more intrinsic frequency domain for DOA estimation. In our Frequency-Domain Convolution (FDC) module, features are transformed via FFT and adaptively filtered by learnable weights, effectively forming a frequency-domain filter that enhances target spatial components while suppressing noise. By directly linking learned features to physical concepts like wavenumber and beam steering, this approach moves beyond the black-box abstraction of purely spatial networks. Consequently, leveraging this physical prior not only enhances model interpretability but also improves estimation accuracy and robustness in challenging low-SNR scenarios. We propose a frequency-domain attention-based Vision Transformer (FViT) for off-grid DOA estimation, with four key contributions:

- Combines coarse classification with fine regression to eliminate grid discretization errors.
- Integrates FFT into the neural network, embedding beamforming physics as learnable inductive bias to enhance model interpretability.
- Proposes the FDC module enabling the class token to better learn frequency-domain features while suppressing noise.
- Uses residual connections and dual-loss training, outperforming both model-based (ESPRIT, R-SBL) and deep learning methods (CNN, DCT-ViT, IQ-ResNet)

under low SNR, limited snapshots, and correlated sources.

## II. SIGNAL MODEL

Consider a uniform linear array (ULA) with  $M$  elements spaced at  $d = \lambda/2$  (where  $\lambda$  is the signal wavelength). For  $K$  far-field narrowband signals  $s_k(t)$  ( $k = 1, 2, \dots, K$ ) incident from  $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_K]$ , the received signal matrix  $\mathbf{X}$  at time  $t$  is modeled as:

$$\mathbf{X}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \quad (1)$$

where  $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_K(t)] \in \mathbb{C}^{K \times T}$  with snapshots  $T$  denotes the signal vector and  $\mathbf{n}(t) \in \mathbb{C}^{M \times T}$  denotes the Additive White Gaussian Noise (AWGN) with zero mean and variance of  $\sigma_n^2$ .  $\mathbf{A} = [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), \dots, \mathbf{a}(\theta_K)] \in \mathbb{C}^{M \times K}$  denotes the manifold matrix, which  $\mathbf{a}(\theta_K)$  can be expressed as

$$\mathbf{a}(\theta_K) = [1, e^{2\pi j \sin \theta_K d/\lambda}, \dots, e^{2\pi j (M-1) \sin \theta_K d/\lambda}]^T \quad (2)$$

where  $[\cdot]^T$  and  $[\cdot]^H$  denote transpose and conjugate transpose, respectively. The theoretical covariance matrix of the received array signal  $\mathbf{X}$  is expressed as:

$$\mathbf{R} = \mathbb{E}[\mathbf{x}(t)\mathbf{x}^H(t)] = \mathbf{A}\mathbf{R}_s\mathbf{A}^H + \sigma_n^2\mathbf{I} \quad (3)$$

where  $\mathbb{E}[\cdot]$  denotes the mathematical expectation, and  $\mathbf{R}_s$  and  $\mathbf{I}$  denote the signal covariance matrix and the identity matrix, respectively. Meanwhile, the sample covariance matrix can be expressed as:

$$\hat{\mathbf{R}} = \frac{1}{T} \sum_{t=0}^T \mathbf{x}(t)\mathbf{x}^H(t) \quad (4)$$

## III. FREQUENCY-DOMAIN ATTENTION ViT FOR DOA ESTIMATION

### A. Data Preprocessing and Labeling

The covariance matrix is decomposed into three real-valued channels (real part, imaginary part, phase) as input  $\mathbf{Y} \in \mathbb{R}^{M \times M \times 3}$ , analogous to RGB image channels, avoiding complex-value handling while preserving spectral-spatial information. DOA estimation is modeled as a dual-label task: classification discretizes the spatial domain  $[\theta_{\min}, \theta_{\max}]$  into

grid  $\mathcal{G} = [\theta_{\min}, \theta_{\min} + \rho, \dots, \theta_{\max} - \rho, \theta_{\max}]$  via uniform grid spacing  $\rho = 1^\circ$ , using one-hot encoding of length  $|\mathcal{G}| = (\theta_{\max} - \theta_{\min})/2 + 1$ ; regression refines the estimate by defining the label as the residual difference between the classified subdomain center and the true angle (typically within  $[-0.5^\circ, 0.5^\circ]$ ). Given the examples  $(-59.8^\circ, -58.2^\circ)$ , the classification label  $\mathcal{L}_1$  and regression label  $\mathcal{L}_2$  are derived as  $\mathcal{L}_1 = [1, 0, 1, \dots, 0]$  and  $\mathcal{L}_2 = [0.2^\circ, -0.2^\circ]$ . Based on the described methodology, the dataset  $\mathcal{D}$  can be structured as  $\mathcal{D} = \{(\mathbf{Y}^{(1)}, \mathcal{L}_1^{(1)}, \mathcal{L}_2^{(1)}), \dots, (\mathbf{Y}^{(D)}, \mathcal{L}_1^{(D)}, \mathcal{L}_2^{(D)})\}$ .

### B. Architecture of FViT

Fig. 1 illustrates the overall network architecture. Our designed network comprises two parallel branches, a classification branch and a regression branch, connected via residual connections. The classification branch primarily consists of FDC modules, specialized for extracting global frequency-domain features. The regression branch incorporates frequency-domain attention to predict the residual between global frequency-domain features and ground-truth labels.

We first apply a point-wise convolution to transform the input feature channels. The Pointwise (PW) convolution uses a kernel size of  $1 \times 1$  with a stride of 1. The input has 3 channels, and the output is 64 channels. This operation can be mathematically represented as:

$$\mathbf{Y}_1 = f_{PWConv}(\mathbf{Y}) \in \mathbb{R}^{B \times C \times 16 \times 16} \quad (5)$$

Subsequently, we feed the output features into the FDC, whose schematic diagram is shown in Fig. 1(b). The specific operations proceed as follows: First, the features pass through a Layer-Norm (LN) layer. Next, a PW-convolution adjusts the channel dimensions. Then, a Fast Fourier transform (FFT) converts the features into the frequency domain. Afterward, the frequency-domain features are multiplied by a learnable scaling factor  $\mathbf{W}$  to amplify task-relevant features while suppressing noise. An inverse Fourier transform (IFFT) then maps the features back to the spatial domain. Finally, a residual connection links the processed features with the original input, mathematically expressed as:

$$\mathbf{Y}_{out} = \mathcal{F}^{-1}(\mathcal{F}(f_{PWConv}(\text{LN}(\mathbf{Y}_{in}))) \otimes \mathbf{W}) + \mathbf{Y}_{in} \quad (6)$$

where  $\mathbf{Y}_{in}$  and  $\mathbf{Y}_{out}$  denote the input feature and output feature, respectively.  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  denote the FFT and IFFT.  $\otimes$  denotes the element-wise multiplication.

After extracting features through multiple FDC modules, we feed them into the regression branch while simultaneously outputting classification results via three fully connected layers with the activation function Gaussian Error Linear Unit (GELU) and dropout layer. The primary challenge of regression tasks stems from the necessity to learn a precise and smooth mapping function within a high-dimensional continuous output space, which constitutes an inherently more complex learning problem compared to learning discrete decision boundaries in classification tasks. We opted for a larger model in the regression branch, specifically, an enhanced frequency-domain ViT. The specific workflow proceeds as follows: First, features

pass through a LN layer and a point-wise convolution layer to expand channel dimensions. Subsequently, the features undergo FFT, followed by concatenation along the channel dimension. The resulting tensor is then flattened along the channel dimension, appended with a learnable class token, and fed into a ViT encoder [14]. After passing through multiple encoder layers, the class token is extracted. This class token is element-wise summed with the output of a PW-convolutional classification branch. The combined feature is then processed through another PW-convolution layer. Finally, analogous to the classification branch, it passes through three fully connected layers to produce K predicted values.

### C. Loss Function

We compute classification loss using binary cross-entropy (BCE) loss and regression loss using Mean Squared Error (MSE) loss, followed by a weighted fusion to obtain the overall loss function, expressed as:

$$Loss = \lambda_1 \ell_{BCE} + \lambda_2 \ell_{MSE} \quad (7)$$

$$\ell_{BCE}(\hat{\mathbf{Y}}_{cls}^i, \mathcal{L}_1^i) = -\frac{1}{|B|} \sum_{n=1}^B [\mathcal{L}_1^i \log(\hat{\mathbf{Y}}^{(n)}) + (1 - \mathcal{L}_1^n) \log(1 - \hat{\mathbf{Y}}^{(n)})] \quad (8)$$

$$\ell_{MSE}(\hat{\mathbf{Y}}_{reg}^i, \mathcal{L}_2^i) = \frac{1}{n} \|\hat{\mathbf{Y}}_{reg}^i - \mathcal{L}_2^i\|_2^2 \quad (9)$$

where  $\lambda_1$  and  $\lambda_2$  denote the weight hyperparameters for the BCE loss function and MSE loss function, respectively.  $B$  represents the batch size, and the  $\|\cdot\|_2$  indicate the L2 norm.

## IV. SIMULATION

### A. Training Scheme

We conduct simulation experiments using an ULA with  $M = 16$  elements. The number of sources is set to  $K = 1, 2$ , and  $3$ , respectively, resulting in  $\mathcal{D} = \binom{121}{1} + \binom{121}{2} + \binom{121}{3} = 295,361$  on-grid angle combinations, which serve as labels for classification results. For the regression branch, we add  $\Delta\theta \sim U(-0.5^\circ, 0.5^\circ)$ , where  $U$  denotes uniform distribution to angle combinations, generating regression labels. Thus, all training samples are off-grid angles. The received signal covariance matrix is constructed via (3) to form the dataset. The SNR ranges from  $-20$  to  $10$  dB with a step size of  $1$  dB, yielding a total of  $|\mathcal{D}| = 295,361 \times 30 = 8,860,830$  samples. The dataset is split into 85% for training and 15% for validation. The network employs the AdamW optimizer with learning rate of 0.001. A cosine annealing schedule adjusts the learning rate dynamically. The cosine annealing schedule follows:

$$\eta_t = \eta_{\min} + \frac{1}{2} (\eta_{\max} - \eta_{\min}) \left( 1 + \cos\left(\frac{T_{cur}}{T_{\max}} \pi\right) \right) \quad (10)$$

where  $\eta_t$  is the current learning rate,  $\eta_{\min}$  and  $\eta_{\max}$  denote the minimum and maximum learning rates,  $T_{cur}$  is the current epoch, and  $T_{\max}$  is the maximum epoch.

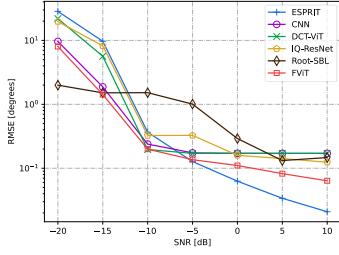


Fig. 2. RMSE vs the SNR ( $T = 1000, K = 2$ )

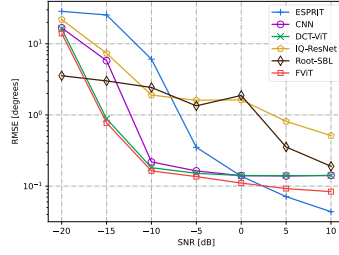


Fig. 3. RMSE vs the SNR ( $T = 1000, K = 3$ )

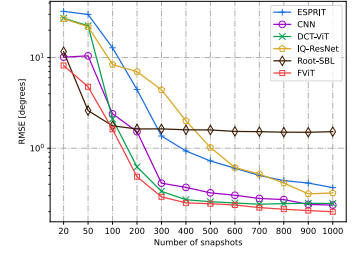


Fig. 4. RMSE vs the snapshots ( $SNR = -10dB, K = 2$ )

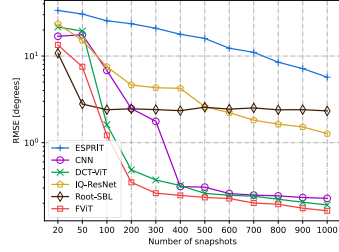


Fig. 5. RMSE vs the snapshots ( $SNR = -10dB, K = 3$ )

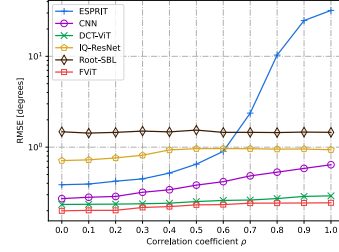


Fig. 6. RMSE vs  $\rho$  ( $SNR = -10dB, T = 1000, K = 3$ )

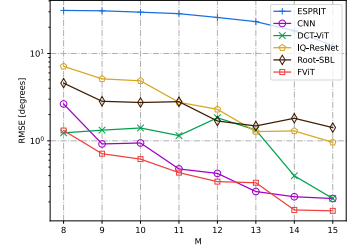


Fig. 7. RMSE vs  $M$  ( $SNR = -10dB, T = 1000, K = 3$ )

## B. Simulation Results

We compare the proposed method with CNN [13], DCT-ViT [16], IQ-ResNet[19], Root-SBL[20] and ESPRIT[2] algorithms under various scenarios including different SNR, snapshots and coherent source conditions. The root mean square error (RMSE) is adopted as the performance metric, defined as:

$$RMSE = \sqrt{\frac{1}{NK} \sum_{k=1}^K \sum_{i=1}^N (\hat{\theta}_k^i - \theta_k^i)^2} \quad (11)$$

where  $N$  denotes the number of Monte-Carlo (MC) experiments and  $\theta_{k,m}$  denotes the estimation of  $\theta_k$  in the  $m$ -th MC experiment.

In the comprehensive evaluation of DOA estimation performance, the proposed FViT method demonstrates consistent superiority across diverse challenging scenarios. Under different SNR conditions with a fixed snapshot count of 1000, FViT achieves the lowest RMSE across the entire tested SNR range (-20 dB to 10 dB) for various source numbers  $K = 2, 3$ , as shown in Fig. 2 and Fig. 3. For instance, with  $SNR = -15dB$ ,  $K = 3$ , FViT attains an RMSE of  $0.77^\circ$ , outperforming DCT-ViT  $0.88^\circ$  and CNN  $5.78^\circ$ . Although slightly trailing Root-SBL at -20 dB, FViT offers significantly lower computational complexity and shows more pronounced performance gains as SNR increases. The method also exhibits strong robustness to snapshot reduction (Fig. 4 and Fig. 5), maintaining superior performance over most comparative methods even at extremely low snapshots (e.g., 20), with continuous improvement as snapshots increase—unlike Root-SBL which shows limited gains. In coherent source scenarios (Fig. 6), FViT's estimation accuracy remains unaffected by source correlation, achieving an RMSE of  $0.24^\circ$  for two coherent sources at  $SNR = -10dB$ , surpassing DCT-ViT  $0.29^\circ$

and CNN  $0.64^\circ$ . Furthermore, in experiments with varying array elements, as shown in Fig. 7, FViT shows robust adaptability, achieving an RMSE of  $0.71^\circ$  with  $M = 9$ , outperforming CNN  $0.92^\circ$  and DCT-ViT  $1.32^\circ$ . These results collectively validate the effectiveness of the hybrid classification-regression scheme and the method's strong generalization capability.

It should be noted that in high-SNR regimes (e.g., above 5 dB), the performance gap between the proposed FViT method and model-based approaches (ESPRIT) narrows. This is expected because: (1) traditional subspace methods achieve near-optimal performance approaching the Cramér-Rao Bound under favorable SNR conditions; (2) the neural network's regression precision has inherent numerical limitations; and (3) the training data distribution emphasizes low-SNR scenarios where data-driven methods provide the greatest advantage. Nevertheless, the proposed method consistently outperforms other deep learning-based on-grid approaches across all tested SNR ranges, demonstrating the effectiveness of the off-grid estimation framework.

## V. CONCLUSION

This paper proposes a novel frequency-domain-based dual-branch ViT network model named FViT for off-grid DOA estimation under low SNR conditions. By extracting frequency-domain features of signals, the model enhances representational capability and interpretability. It achieves off-grid estimation through residual connections that integrate features from both the classification branch and regression branch. Meanwhile, the dimension-preserving property of the network ensures stackable capability, enabling seamless deep extension to further improve network performance. Simulation results demonstrate that the proposed FViT method outperforms other existing approaches under low SNR conditions.

## REFERENCES

- [1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986, doi: 10.1109/TAP.1986.1143830.
- [2] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 7, pp. 984–995, July 1989, doi: 10.1109/29.32276.
- [3] X. Wu, X. Yang, X. Jia, and F. Tian, "A Gridless DOA Estimation Method Based on Convolutional Neural Network With Toeplitz Prior," *IEEE Signal Process. Lett.*, vol. 29, pp. 1247–1251, 2022, doi: 10.1109/LSP.2022.3176211.
- [4] S. Zhou, X. Ma, and W. Sheng, "ESPRIT for Time-Modulated Array With Unidirectional Phase Center Motion," *IEEE Signal Process. Lett.*, vol. 32, pp. 3117–3121, 2025, doi: 10.1109/LSP.2025.3585817.
- [5] X. Shen, Z. Zhuang, H. Wang, and F. Shu, "An Effective Method for Attributed Scattering Center Extraction Based on an Improved ESPRIT Algorithm," *IEEE Trans. Antennas Propag.*, vol. 73, no. 3, pp. 1618–1629, Mar. 2025, doi: 10.1109/TAP.2024.3502907.
- [6] J. Bai *et al.*, "A Novel-Deep-Neural-Network-Architecture-Based GAN-DRANet for DOA Sensing With an Enhanced Performance in Low SNR," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 22610–22622, July 2025, doi: 10.1109/IJOT.2025.3549438.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv: arXiv:1810.04805*. Accessed: Nov. 05, 2024. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [8] Y. Ma, Y. Zeng, and S. Sun, "A Deep Learning Based Super Resolution DoA Estimator With Single Snapshot MIMO Radar Data," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4142–4155, Apr. 2022, doi: 10.1109/TVT.2022.3151674.
- [9] J. P. Merkofer, G. Revach, N. Shlezinger, T. Routtenberg, and R. J. G. van Sloun, "DA-MUSIC: Data-Driven DoA Estimation via Deep Augmented MUSIC Algorithm," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 2771–2785, Feb. 2024, doi: 10.1109/TVT.2023.3320360.
- [10] S. Zheng *et al.*, "Deep Learning-Based DOA Estimation," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 3, pp. 819–835, June 2024, doi: 10.1109/TCCN.2024.3360527.
- [11] X. Zhao, J. Atli Benediktsson, Y. Yang, K.-S. Chen, and M. Örn Úlfarsson, "Exploring Transformer-Based Direction-of-Arrival Estimation Over Sea Surface: A BERT Approach With Physics-Based Loss Function," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, doi: 10.1109/TGRS.2024.3440224.
- [12] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," Jan. 23, 2020, *arXiv: arXiv:2001.08361*. doi: 10.48550/arXiv.2001.08361.
- [13] G. K. Papageorgiou, M. Sellathurai, and Y. C. Eldar, "Deep Networks for Direction-of-Arrival Estimation in Low SNR," *IEEE Trans. Signal Process.*, vol. 69, pp. 3714–3729, 2021, doi: 10.1109/TSP.2021.3089927.
- [14] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 03, 2021, *arXiv: arXiv:2010.11929*. doi: 10.48550/arXiv.2010.11929.
- [15] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- [16] Y. Guo, Z. Zhang, and Y. Huang, "Dual Class Token Vision Transformer for Direction of Arrival Estimation in Low SNR," *IEEE Signal Process. Lett.*, vol. 31, pp. 76–80, 2024, doi: 10.1109/LSP.2023.3342628.
- [17] Y. Xie, A. Liu, X. Lu, and D. Chong, "Hybrid Multi-Class Token Vision Transformer Convolutional Network for DOA Estimation," *IEEE Signal Process. Lett.*, vol. 32, pp. 2279–2283, 2025, doi: 10.1109/LSP.2025.3573949.
- [18] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, June 2023, pp. 5886–5895. doi: 10.1109/CVPR52729.2023.00570.
- [19] S. Zheng *et al.*, "Deep Learning-Based DOA Estimation," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 3, pp. 819–835, June 2024, doi: 10.1109/TCCN.2024.3360527.
- [20] J. Dai, X. Bao, W. Xu, and C. Chang, "Root Sparse Bayesian Learning for Off-Grid DOA Estimation," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 46–50, Jan. 2017, doi: 10.1109/LSP.2016.2636319.
- [21] B. A. Johnson, Y. I. Abramovich, and X. Mestre, "MUSIC, G-MUSIC, and Maximum-Likelihood Performance Breakdown," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3944–3958, Aug. 2008, doi: 10.1109/TSP.2008.921729.
- [22] F. Li and R. J. Vaccaro, "Sensitivity analysis of DOA estimation algorithms to sensor errors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 708–717, July 1992, doi: 10.1109/7.256292.
- [23] M. Wax and J. Sheinvald, "Direction finding of coherent signals via spatial smoothing for uniform circular arrays," *IEEE Transactions on Antennas and Propagation*, vol. 42, no. 5, pp. 613–620, May 1994, doi: 10.1109/8.299559.