

RobWDM: Hardware-Accelerated CNN Meeting Wavenumber-Domain Method for Robust Near-Field Sensing

Anouar Nechi, Mohammadreza Fouji, Rainer Buchty, and Mladen Berekovic
Institute of Computer Engineering, University of Lübeck, Germany

Abstract—Sixth-generation (6G) communication systems leveraging Extremely Large Aperture Arrays (ELAAs) are increasingly adopting near-field sensing (NFS) for joint angle and distance estimation. However, extracting two-dimensional positional information introduces substantial computational overhead. Conventional Multiple Signal Classification (MUSIC) methods suffer from high latency due to exhaustive grid searches, while recent wavenumber-domain approaches incur large memory costs and provide limited validation under hardware impairments. To address these challenges, we propose RobWDM, a hardware-accelerated wavenumber-domain framework for robust near-field sensing. We further introduce LightCNN, a hardware-aware ultra-compact CNN optimized for low-latency FPGA/edge inference in wavenumber-domain sensing pipelines. FPGA deployment achieves inference latencies as low as 51 μ s, providing a 70 $\times$ speedup over CPU execution while maintaining competitive localization accuracy and improved energy efficiency and robustness to hardware impairments, maintaining bounded localization error under moderate mismatch conditions. These results demonstrate the practical viability of FPGA-accelerated near-field sensing for future 6G systems.

Index Terms—Deep learning, near-field sensing (NFS), target localization, wavenumber-domain, robustness, field-programmable gate arrays (FPGA).

I. INTRODUCTION

High-precision target localization is a cornerstone of emerging sixth-generation (6G) wireless systems. The massive integration of Extremely Large Aperture Arrays (ELAAs) has fundamentally altered the physical landscape of signal propagation, driving a shift from traditional far-field assumptions to full-dimensional near-field sensing (NFS) [1], [2]. As array dimensions scale, electromagnetic waves depart from the classical planar-wave approximation and enter the near-field regime, where signals exhibit spherical wavefronts and spatially varying phase profiles across the array aperture.

By exploiting these unique phase variations, NFS enables the capture of the target’s angle and radial distance from a single observation [3], [4]. While this enables super-resolution localization, extracting such two-dimensional positional data poses severe computational challenges. Conventional spatial spectrum estimation techniques are inherently constrained by the complex coupling between distance and angle, resulting in significant bottlenecks that threaten the real-time viability of sensing solutions [5]. Consequently, achieving high-precision localization within the strict latency and reliability constraints of 6G infrastructure requires a comprehensive re-engineering

of current models to support robust operation at the network edge.

A. Related Works & Motivations

Traditional spatial spectrum estimation, especially the two-dimensional (2D) Multiple Signal Classification (MUSIC) algorithm, has been a key approach for near-field parameter estimation in array signal processing [6], [3]. However, 2D MUSIC has high computational complexity of $\mathcal{O}(M^3 + M^2G_1G_2)$, where M is antenna count and G is grids for angle and distance [5]. To address this, researchers have proposed “angle-then-distance” methods that break 2D searches into two 1D searches, lowering complexity to $\mathcal{O}(M^3 + M^2(G_1 + G_2))$, where G_1 and G_2 denote the number of search grid points for angle and distance, respectively [7], [8]. Other improvements use subspace-fitting MUSIC with the Fresnel approximation to separate angle and distance, greatly shrinking the search space dimension [9].

Beyond computational constraints, classical grid-based approaches exhibit severe vulnerabilities to real-world conditions, degrading significantly under hardware impairments or a lack of observation data. To overcome these limitations, recent studies propose AI-aided subspace methods that utilize Deep Neural Networks (DNNs) to learn robust surrogate covariance matrices from data. For instance, NF-SubspaceNet [10] augments MUSIC by introducing a differentiable spectrum loss that directly optimizes the surrogate matrix, thereby maintaining orthogonality between the noise subspace and the true target despite flawed hardware. Additionally, order-aware training schemes integrate specific statistical criteria to help the network distinguish signal from noise. While NF-SubspaceNet addresses hardware impairments, it still relies on heavy 2D grid searches. To resolve hardware errors and computational bottlenecks, cascaded hybrid architectures like Deep-learning-aided Cascaded Differentiable MUSIC (DCD-MUSIC) [11] decouple angle and range estimation. DCD-MUSIC uses an autoencoder to recast near-field signals into a far-field surrogate representation for efficient angle estimation via ESPRIT [12]. A secondary DNN then generates a surrogate covariance for 1D MUSIC distance estimation. This cascaded approach eliminates the exhaustive 2D search and enables zero-shot generalization for localizing mixed near-field and far-field targets.

Alternatively, researchers have bypassed spatial subspace methods entirely by mapping channels into the wavenumber domain, where power-concentrated regions are sparse and directly related to target positions. Here, bi-directional convolutional neural networks (Bi-CNNs) learn these wavenumber patterns to achieve gridless positioning with a significantly reduced complexity of $\mathcal{O}(M^2)$ [13].

Despite these advancements, a critical gap remains for practical edge deployment. Both hybrid AI-aided methods and wavenumber-domain Bi-CNNs retain large parameter footprints that exceed on-chip memory limits. Furthermore, many learning-based models assume ideal conditions and remain susceptible to practical radio-frequency (RF) impairments, such as phase mismatches. These challenges motivate the development of an ultra-compact, hardware-accelerated framework capable of robust, sub-millisecond localization in a non-ideal hardware setup.

B. Paper Contributions

Motivated by the stringent latency and reliability requirements of 6G networks, this paper proposes RobWDM, a hardware-aware framework for near-field target localization. By jointly optimizing the neural network architecture and FPGA execution pipeline, the proposed framework enables gridless wavenumber-domain sensing with sub-millisecond inference latency and improved robustness to hardware impairments. The main contributions of this work are summarized as follows:

- We propose LightCNN, a lightweight CNN architecture for low-latency near-field localization that achieves sub-millisecond inference latency while maintaining competitive positioning accuracy. Compared with conventional 2D MUSIC and baseline Bi-CNN methods, LightCNN provides substantially lower computational complexity and faster inference.
- We implement LightCNN on an FPGA platform and demonstrate energy-efficient real-time inference suitable for latency-constrained 6G edge deployments.
- We evaluate LightCNN's robustness under hardware impairments, including phase and distance mismatches across multiple array configurations. Experimental results show that LightCNN maintains lower RMSE and improved localization stability compared with baseline approaches under increasing hardware impairments.

II. SYSTEM MODEL

Following the system model established in [13], we consider a full-duplex multiple-input single-output (MISO) system with a base station (BS) and a single antenna target. As illustrated in Fig. 1, the BS employs a uniform linear array (ULA) with $M = 2\bar{M} + 1$ antennas, where $\bar{M}$ denotes the number of elements on each side of the central antenna. The array is aligned along the x-axis, with inter-element spacing $d = 0.5\lambda$ and total aperture $D = (M - 1)d$. The coordinates of the m -th antenna element and the target can be expressed as $x_m =$

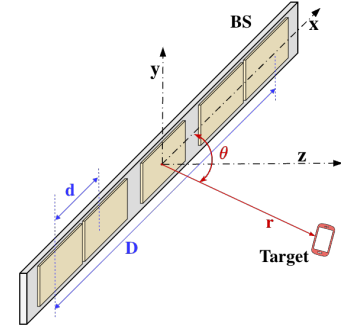


Fig. 1: Near-field MISO system model.

$[x_m^x, x_m^y, x_m^z]$ and $\mathbf{r} = [r^x, r^y, r^z]$ in the x-z plane, respectively.

The target is assumed to be in the radiating near-field region of the BS array, where the distance r satisfies $r < \frac{2D^2}{\lambda}$. Because the planar-wave assumption does not hold in this region [4], a near-field channel model is adopted to account for spherical wavefronts. A single narrow frequency band with bandwidth B is employed for localization. The near-field array response vector is given by:

$$\mathbf{a}(\mathbf{r}) = \left[e^{-j\frac{2\pi}{\lambda}\|\mathbf{r}-\mathbf{x}_{-\bar{M}}\|_2}, \dots, e^{-j\frac{2\pi}{\lambda}\|\mathbf{r}-\mathbf{x}_{+\bar{M}}\|_2} \right]^T \quad (1)$$

where $\|\cdot\|_2$ denotes the two-norm of a vector. Consequently, the line-of-sight (LoS) channel matrix $\mathbf{H} \in \mathbb{C}^{M \times M}$ is modeled as:

$$\mathbf{H} = \beta \mathbf{a}(\mathbf{r}) \mathbf{a}^T(\mathbf{r}) \quad (2)$$

where β denotes the complex channel gain. This gain is defined as $\beta \triangleq \zeta_{\text{pathloss}}(f, 2r) G_t G_r$, where G_t and G_r represent the transmit and receive array gains, respectively. The pathloss function is given by $\zeta_{\text{pathloss}}(f, d) = \frac{c}{4\pi f d}$, where c is the speed of light. Given that the target is located within the radiating near-field region, the pathloss variances among individual antenna elements are negligible and therefore replaced by the pathloss of the center link [3].

Let $s[n] \in \mathbb{C}$ denote the probing signal at the BS during the n -th coherent time block. The transmitted signal is formulated as $\mathbf{x}[n] = \sqrt{P} \mathbf{w} s[n]$, where P represents the BS's transmit power and $\mathbf{w} \in \mathbb{C}^{M \times 1}$ is the precoding vector subject to a unit power constraint ($\|\mathbf{w}^H \mathbf{w}\| = 1$). Following propagation through the round-trip channel, the corresponding echo signal received at the base station is defined as

$$\mathbf{y}[n] = \mathbf{H} \mathbf{x}[n] + \mathbf{z}[n] \quad (3)$$

where $\mathbf{z}[n] \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ denotes the complex Gaussian noise with zero mean and covariance $\sigma^2 \mathbf{I}$.

At high-frequency bands, wireless channels are predominantly LoS-dominated due to severe scattering losses of non-line-of-sight (NLoS) components [14]. The target's position is embedded within the LoS channel, its representation in the space-domain channel $\mathbf{H}$ is highly spread and non-explicit, which typically entails complex signal processing for extraction. To circumvent high computational complexity, the

space-domain channel representation is transformed into the wavenumber domain, where the target's position can be more clearly observed.

III. METHODOLOGY

To circumvent the high complexity of grid-based spatial spectrum searches, the RobWDM framework transforms space-domain channel data into a low-overhead, gridless position estimate using a two-stage processing pipeline. First, the raw channel matrix is mapped into the wavenumber domain via a discrete semi-unitary transformation. This step concentrates diffuse spherical wavefront variations into localized, high-energy multi-channel features that explicitly encode target angle and distance. Second, these structured observations are mapped directly to a 2D coordinate estimate by LightCNN, a hardware-accelerated CNN optimized for low-latency FPGA deployment.

A. Wavenumber-Domain Transformation

The fundamental concept behind the wavenumber-domain transformation is to approximate near-field spherical wavefronts using a combination of planar wavefronts, yielding a structure identical to that of a 4D Fourier transform. In contrast to the traditional wavenumber-domain approach outlined in [15], the transceivers in this study are co-located due to the integration of a full-duplex antenna architecture. Accordingly, the 4D Fourier transform reduces to a 2D version, enabling the entry in the i -th column and j -th row of matrix H to be written as:

$$[H]_{i,j} = \frac{1}{4\pi^2} \iint_{\mathcal{D}_k} a(k, x_m) h_a(k_x, k_y) a^*(k, x_m) dk_x dk_y \quad (4)$$

The region of integration, $\mathcal{D}_k$, is defined as $\mathcal{D}_k \triangleq \{(k_x, k_y) \in \mathbb{R}^2 : k_x^2 + k_y^2 \leq k_0^2\}$. Within this region, the wave vector is defined as $k \triangleq [k_x, k_y, \gamma(k_x, k_y)]^T$, where $\gamma(a, b) \triangleq \sqrt{k_0^2 - a^2 - b^2}$. Consequently, $a(k, x_m) = \exp\{-jk^T x_m\}$ represents the array response, and $h_a(k_x, k_y)$ denotes the coupling coefficient between the transmit and receive array responses.

To ensure the computational tractability of (4) for low-latency hardware implementation, the continuous integral over the region $[-k_0, +k_0]$ is evaluated numerically. This discretization is performed by applying a uniform sampling interval of $2\pi/D$. Furthermore, because the entire system is confined to the x-z plane, the y-coordinate is omitted, resulting in a discretized region $\mathcal{G}_k \triangleq \{\epsilon \in \mathbb{Z} : -\lceil Dk_0/2\pi \rceil \leq \epsilon \leq \lfloor Dk_0/2\pi \rfloor\}$. As a result, the integral is converted into a summation that is equivalent to matrix multiplication:

$$H \cong M \sum_{i=1}^{|\mathcal{G}_k|} \sum_{j=1}^{|\mathcal{G}_k|} a(i) [H_a]_{i,j} a^*(j) = M A H_a A^H \quad (5)$$

Here H_a denotes the wavenumber-domain channel representation. Consequently, the array response can be written as

$$a(i) = \frac{1}{\sqrt{M}} \left[e^{-j \frac{2\pi i}{D} x_{-\tilde{M}}^x}, \dots, e^{-j \frac{2\pi i}{D} x_{+\tilde{M}}^x} \right] \quad (6)$$

The accuracy of the discretization approximation is highly robust when the normalized aperture size D exceeds several wavelengths λ [15]. Given that the wavenumber-domain transformation matrix (WTM), $A \triangleq [a(1), \dots, a(|\mathcal{G}_k|)]$, is semi-unitary ($A^H A = \mathbf{I}$), the wavenumber-domain channel is derived as:

$$H_a = \frac{1}{M} A^H H A \quad (7)$$

The application of this wavenumber-domain transformation serves as a critical preprocessing step, mitigating spatial-domain ambiguities and facilitating direct extraction of the target's spatial features. By sampling the directional wave vector k , the incident signals are mapped to localized high-energy clusters. In the near-field context, the inherent wavefront curvature encodes both radial distance and angular direction within these clusters, thereby generating a highly structured feature space optimized for the proposed CNN model.

Employing this transformation as a preprocessing stage reveals the target's location information embedded in the near-field channels. Because H_a is obtained by sampling the directional wavenumber k , its power-concentrated regions correlate directly with the impinging signal directions. Due to the near-field phenomenon, distance information is inherently encoded within these directions, enabling simultaneous angle and distance estimation.

To extract the target's location from this representation, the BS transmits a probing signal s using a unit power transmit beamformer $w = w'/\|w'\|_2$, where $w' = A\mathbf{1}$ and $\mathbf{1} \in \mathbb{R}^{M \times 1}$ is an all-one vector. Utilizing a single snapshot for localization, the received echo signal y is processed via matrix multiplication by A^H and divided by s to yield the observation vector $\tilde{o}$:

$$\tilde{o} = \frac{A^H y}{s} = \frac{\sqrt{P} A^H A H_a A^H A \mathbf{1}}{\|w'\|_2} + \frac{A^H z}{s} \quad (8)$$

$$\tilde{o} \stackrel{(a)}{=} c H_a \mathbf{1} + \tilde{z}$$

Step (a) is achieved by leveraging the semi-unitary property of the transformation matrix, where $c = \sqrt{P}/\|w'\|_2$ is a constant and $\tilde{z} \triangleq A^H z/s$ represents the processed noise. Unlike [13], we construct a three-channel input by jointly exploiting the magnitude and phase of $\tilde{o}$. To suppress noise and isolate high-confidence signal regions, a binary mask b is defined as

$$b = \mathbb{1} \left(\frac{|\tilde{o}|}{\max(|\tilde{o}|)} > 0.5 \right) \quad (9)$$

where $|\cdot|$ denotes the absolute value operator and $\mathbb{1}(\cdot)$ is the indicator function. The final preprocessed input o is formed by stacking these features:

$$o = \left(\frac{|\tilde{o}|}{\max(|\tilde{o}|)} \odot b, \text{Re} \left\{ \frac{\tilde{o} \odot b}{|\tilde{o}|} \right\}, \text{Im} \left\{ \frac{\tilde{o} \odot b}{|\tilde{o}|} \right\} \right) \in \mathbb{R}^{3 \times N} \quad (10)$$

where $\odot$ denotes the Hadamard product. The first channel represents the normalized magnitude, while the second and third channels correspond to the cosine and sine components of the phase on the masked support, respectively.

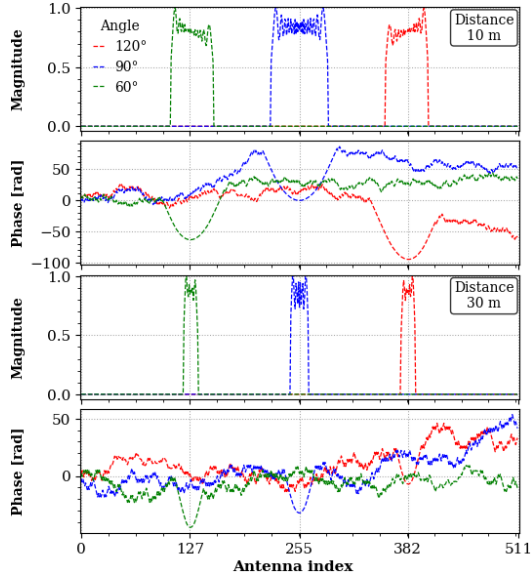


Fig. 2: Normalized magnitude and unwrapped phase of the observations for different targets

Fig. 2 illustrates representative preprocessed observations of magnitude for three distinct target positions along the array, at distances of 10 m and 30 m. Each observation exhibits a well-localized, power-concentrated region whose position along the antenna index axis encodes the target angle, while the spread and amplitude of the region vary with distance, reflecting the curvature of the near-field spherical wavefront. Notably, the binary mask effectively suppresses the noise floor outside the signal support, retaining only the high-confidence region regardless of the target’s position or distance, which is critical for robust CNN inference under varying channel conditions. This multi-channel representation preserves the critical phase variations inherent to near-field spherical wavefronts, providing a rich feature set for the subsequent LightCNN architecture.

B. Neural Topology Search

To obtain a compact CNN suitable for deployment on resource-constrained FPGA hardware, we performed an exhaustive architecture search over a predefined candidate space. The search objective is to identify the topology that minimizes the localization distance RMSE while strictly constraining the number of trainable parameters to below 10K, ensuring the model fits within the available on-chip memory of the FPGA.

All candidate architectures share the same general structure, inspired by the Bi-CNN topology of [13]. Two key distinctions are introduced: the bidirectional structure is replaced by a strictly sequential, one-dimensional convolutional pipeline to reduce the parameter footprint, and the Gaussian Error Linear Unit (GeLU) activation is replaced with Rectified Linear Unit (ReLU) across all intermediate layers. The latter choice is motivated by hardware efficiency, as GeLU requires evaluating the Gaussian cumulative distribution function, an operation that is costly to implement in fixed-point FPGA arithmetic,

whereas ReLU reduces to a simple thresholding operation with negligible impact on localization accuracy at this model scale.

TABLE I: LightCNN Layer-wise Output Shapes and Parameter Counts for $M=255$ and $M=511$

Layers	Output Shape		Parameters	
	$M=255$	$M=511$	$M=255$	$M=511$
Conv1D + ReLU	(64, 255)	(64, 511)	960	960
MaxPool1D	(64, 63)	(64, 127)	0	0
Conv1D + ReLU	(12, 63)	(12, 127)	2,304	2,304
MaxPool1D	(12, 15)	(12, 31)	0	0
Flatten	(180)	(372)	0	0
Linear + ReLU	(8)	(8)	1,440	2,976
Linear ¹	(2)	(2)	18	18
LightCNN parameters			4,722	6,258
Bi-CNN parameters			2,092,586	2,616,874

Each model takes the three-channel wavenumber-domain observation defined in (10) as input. The search space was defined as follows. The number of output channels in the first and second convolutional layers was varied over $\{8, 16, 32, 64\}$ and $\{12, 24, 32, 64\}$, respectively. Two subsets of fully connected (FC) configurations were considered: a single FC layer with units drawn from $\{8, 16, 32, 64\}$, and a two-layer configuration where the first and second layers were varied over $\{8, 16, 32, 64\}$ and $\{8, 16\}$, respectively. The latter subset deliberately excludes larger FC-layer sizes, as these configurations would exceed the 10K-parameter budget for the considered ranges of convolutional channels. This yields a total candidate pool of 192 architectures. Each candidate was trained for up to 200 epochs with an early stopping mechanism, under identical conditions to ensure a fair comparison. Among all architectures satisfying the parameter constraint, the model achieving the lowest distance RMSE was selected as the final architecture, referred to hereafter as LightCNN.

The resulting architecture is detailed in Table I. The two array configurations share identical parameters across all layers except the fully connected layer, where the flattened dimension differs due to the different input lengths. This results in a total of only 4,722 and 6,258 trainable parameters for $M = 255$ and $M = 511$, respectively. The baseline Bi-CNN model requires 2,092,586 and 2,616,874 parameters for the same array sizes. By satisfying the 10K parameter budget, LightCNN achieves a more than 400 $\times$ reduction in memory footprint compared to the baseline, ensuring it easily fits within the target FPGA’s on-chip memory constraints.

IV. EXPERIMENTAL RESULTS

This section evaluates the proposed RobWDM framework across two complementary dimensions: positioning accuracy under realistic hardware impairments, and inference efficiency on embedded hardware. All simulations share a common physical-layer configuration and dataset, ensuring a fair comparison across all evaluated methods. Furthermore, quantitative results for localization performance, robustness, and FPGA acceleration outcomes are detailed and discussed.

¹Bias parameters are disabled for all convolutional and intermediate linear layers, and are only enabled in the final output Linear layer.

A. Experimental Setup

Simulations employ a full-duplex MISO configuration with a 28 GHz BS and a single-antenna target. We evaluate ULA sizes of $M \in \{255, 511\}$ with half-wavelength spacing. Array gains are $G_r = 10^{0.5}$ and $G_t = 10^{1.5}$. The thermal noise floor is set to -75 dBm.

TABLE II: Dataset generation and training parameters

Parameter	Value
Dataset Generation	
Angular range [rad]	$[-\pi/4, \pi/4]$
Distance range [m]	[8, 35]
Resolution	0.01
Hardware impairments	
Phase mismatch (δ_ϕ) [rad]	$\{0.05, 0.10, 0.20\}$
Distance mismatch (δ_d) [m]	$\{0.01, 0.02, 0.05\}$
Training parameters	
Optimizer	AdamW
Loss	RMSE
Learning rate	10^{-3}
Weights decay	10^{-5}

As detailed in Table II, the dataset is generated by exhaustively sampling the target space over the angular and distance ranges at a resolution of 0.01. To evaluate robustness under non-ideal conditions, hardware impairments, including phase δ_ϕ and distance mismatches, are introduced independently at test time. The proposed CNN models are trained for up to 200 epochs, with an early stopping patience of 10 epochs to prevent overfitting.

Localization accuracy is quantified by the RMSE in meters, averaged over the test set. We consider two baselines: the 2D MUSIC algorithm and the Bi-CNN model in [13]. The FPGA implementation targets the AMD Xilinx ZCU104 platform. The LightCNN inference datapath is implemented in fixed-point arithmetic and synthesized at 100 MHz using Vivado 2020.1. We extract resource utilization, throughput, and power metrics from post-implementation reports.

B. Positioning Accuracy & Robustness

The localization accuracy of LightCNN is compared to 2D MUSIC and Bi-CNN for both BS sizes on CPU, as shown in Table III. In ideal conditions, LightCNN achieves a distance RMSE of 0.09 m for $M = 255$ and 0.07 m for $M = 511$. This is 6.9 $\times$ and 3.4 $\times$ better than Bi-CNN. 2D MUSIC, even at finer grid 10^3 , shows an RMSE of 0.016 m for $M = 255$ and 0.015 m for $M = 511$. However, its latency is much higher: 11.2s and 23s, three to four orders of magnitude slower than LightCNN. On a coarser grid of 10^2 , MUSIC accuracy drops, with RMSEs of 1.61 m and 1.11 m, and run times of 0.94 s and 1.2 s. This makes MUSIC impractical for low-latency needs. LightCNN, in contrast, finishes in 3.7 ms and 5.2 ms for $M = 255$ and $M = 511$. This is a 2.6 $\times$ speedup over Bi-CNN, and also gives better positioning. These results show that the compact one-dimensional convolutional design, found via neural topology search, achieves higher accuracy and faster speed. It outperforms both the exhaustive spectral method and the larger CNN baseline.

The robustness of LightCNN to hardware impairments is evaluated in Fig. 3, where the positioning RMSE is shown

TABLE III: Positioning accuracy and CPU run time comparison between 2D MUSIC, Bi-CNN, and LightCNN

M	Method	RMSE (m)	Run Time (ms)
255	MUSIC	10^2	1.61
		10^3	0.016
	Bi-CNN	0.62	9.5
	LightCNN	0.09	3.7
511	MUSIC	10^2	1.11
		10^3	0.015
	Bi-CNN	0.24	10.2
	LightCNN	0.07	5.2

as a function of phase δ_ϕ and distance δ_d mismatches. Under ideal conditions, LightCNN reports RMSEs of 0.09 m and 0.07 m for $M = 255$ and $M = 511$, respectively. These results are consistent with values in Table III. As mismatch severity increases, LightCNN shows a more stable degradation profile than Bi-CNN across both BS sizes.

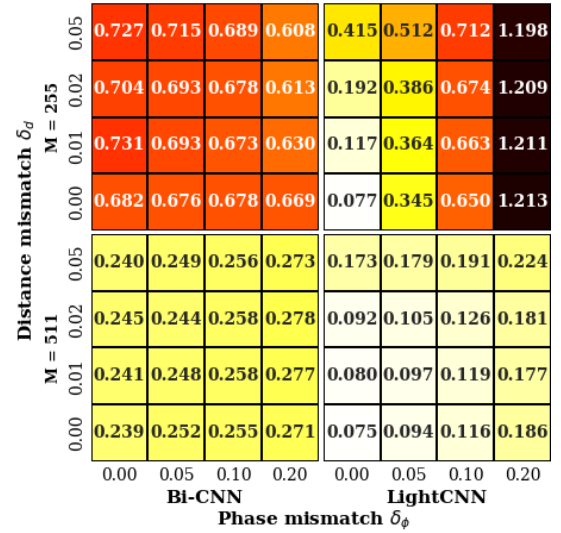


Fig. 3: Heatmap of positioning RMSE (m) between Bi-CNN and LightCNN under varying distance and phase mismatches for different array sizes

For $M = 255$, LightCNN demonstrates superior precision under near-ideal conditions. It reaches a peak RMSE of 1.213 m under extreme phase mismatch ($\delta_\phi = 0.20$ rad). In contrast, Bi-CNN maintains a flatter response across the impairment range but has a consistently higher error floor. RMSE values remain above 0.60 m even at minimal mismatch levels. For $M = 511$, the contrast is more evident. LightCNN remains below 0.19 m across most tested mismatch combinations, with a maximum of 0.224 m. Bi-CNN, under similar conditions, reaches 0.278 m. This is a 19.4% reduction in worst-case RMSE for the larger array. LightCNN's robustness advantage becomes more apparent as the aperture grows.

These findings arise from two complementary factors. First, the three-channel input representation in (10) encodes both the magnitude and phase of the wavenumber-domain observation, providing the network with sufficient information to implicitly compensate for perturbations in the array response. Second, LightCNN's compact parameter count limits overfitting on

ideal training data, leading to a model that generalizes more effectively to non-ideal conditions at test time. Taken together, these properties establish LightCNN as a viable localization engine for ELAA deployments, where unavoidable calibration errors and hardware aging must be considered.

C. Hardware Implementation

The FPGA implementation results for LightCNN on the AMD Xilinx ZCU104 platform, detailed in Table IV, demonstrate an optimized balance between resource utilization, inference throughput, and energy efficiency. Synthesized at a clock frequency of 100 MHz, the dedicated dataflow accelerators achieve hardware latencies of approximately 51 μ s for $M = 255$ and 96 μ s for $M = 511$. These execution times represent a substantial acceleration, exceeding 70 $\times$, relative to the CPU-based inference reported in Table III, while outperforming the 2D MUSIC baseline. Consequently, LightCNN facilitates real-time, sub-millisecond localization, effectively supporting applications with rapid target motion or high computational load.

TABLE IV: Implementation Results and Resource Utilization on ZCU104

Metric	$M = 255$	$M = 511$
Frequency [MHz]	100	100
Resource Usage		
LUT	62 868 (27.3%)	65 979 (28.6%)
FF	61 075 (13.2%)	78 568 (17.0 %)
DSP	419 (24.2%)	442 (25.6 %)
BRAM	194 (62.2%)	242 (77.6 %)
Performance		
Throughput [Inf/s]	19.6 K	10.4 K
Dynamic Power [W]	0.157	0.192
Efficiency [Inf/J]	124.8 K	54.3 K

The implementation maintains a modest hardware footprint, with total resource consumption remaining well within the FPGA's available capacity. BRAM constitutes the primary resource bottleneck for both array configurations, driven by the on-chip storage requirements for quantized weights and intermediate feature maps. Despite a marginal increase in dynamic power consumption for the larger $M = 511$ array, due to expanded BRAM and DSP utilization, the overall power requirements remain strictly within the normal ranges for such a design. High energy efficiency in both cases further confirms that LightCNN meets the strict latency and energy requirements of next-generation ELAA sensing.

Compared with the Bi-CNN baseline, LightCNN offers distinct architectural advantages for hardware deployment. The increased parameter count and structural complexity of Bi-CNN would necessitate substantially higher BRAM allocation, resulting in increased latency and reduced throughput. Furthermore, the strategic substitution of GeLU with ReLU activations eliminates the need for resource-intensive lookup tables or complex polynomial approximations in fixed-point arithmetic. Collectively, these design decisions, comprising a compact topology, hardware-friendly activations, and fixed-point quantization, ensure the RobWDM framework's deployment readiness for practical sensing environments.

V. CONCLUSION

This paper presented RobWDM, a hardware-aware framework that combines wavenumber-domain preprocessing with the proposed ultra-compact LightCNN to enable robust, real-time near-field sensing for future 6G ELAA systems. Experimental results demonstrated that LightCNN achieves low localization error and improved robustness to hardware impairments compared with conventional 2D MUSIC and baseline Bi-CNN methods. FPGA deployment achieved inference latencies as low as 51 μ s, corresponding to a 70 $\times$ speedup over CPU execution while maintaining energy-efficient operation. These results demonstrate the practical feasibility of low-latency, FPGA-accelerated near-field sensing in realistic non-ideal hardware environments.

REFERENCES

- [1] Z. Zheng, M. Fu, W.-Q. Wang, S. Zhang, and Y. Liao, "Localization of mixed near-field and far-field sources using symmetric double-nested arrays," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 11, pp. 7059–7070, 2019.
- [2] Q. Yang, A. Guerra, F. Guidi, N. Shlezinger, H. Zhang, D. Dardari, B. Wang, and Y. C. Eldar, "Beam focusing for near-field multi-user localization," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 8, pp. 12 259–12 273, 2025.
- [3] Z. Wang, X. Mu, and Y. Liu, "Near-field integrated sensing and communications," *IEEE Communications Letters*, vol. 27, no. 8, pp. 2048–2052, 2023.
- [4] Y. Liu, Z. Wang, J. Xu, C. Ouyang, X. Mu, and R. Schober, "Near-field communications: A tutorial review," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 1999–2049, 2023.
- [5] Y.-D. Huang and M. Barkat, "Near-field multiple source localization by passive sensor array," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 7, pp. 968–975, 1991.
- [6] X. Zhang, W. Chen, W. Zheng, Z. Xia, and Y. Wang, "Localization of near-field sources: A reduced-dimension music algorithm," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1422–1425, 2018.
- [7] H. Li, Z. Wang, X. Mu, P. Zhiwen, and Y. Liu, "Near-field integrated sensing, positioning, and communication: A downlink and uplink framework," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 9, pp. 2196–2212, 2024.
- [8] K. Qu, S. Guo, J. Ye, and H. Zhao, "Two-stage beamspace music-based near-field channel estimation for hybrid xl-mimo," *IEEE Communications Letters*, vol. 28, no. 8, pp. 1949–1953, 2024.
- [9] R. Zhang, Z. Wang, Z. Wei, Y. Liu, Z. Xiong, and Z. Feng, "Subspace fitting approach for wideband near-field localization," 2025. [Online]. Available: <https://arxiv.org/abs/2508.04169>
- [10] A. Gast, L. Le Magoarou, and N. Shlezinger, "Near-field localization via ai-aided subspace methods," *IEEE Transactions on Communications*, vol. 74, pp. 6117–6132, 2026.
- [11] —, "Dcd-music: Deep-learning-aided cascaded differentiable music algorithm for near-field localization of multiple sources," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [12] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.
- [13] H. Jiang, Z. Wang, and Y. Liu, "Near-field sensing: A low-complexity wavenumber-domain positioning method," *IEEE Wireless Communications Letters*, vol. 14, no. 4, pp. 994–998, 2025.
- [14] A. Tang, J.-B. Wang, Y. Pan, W. Zhang, Y. Chen, H. Yu, and R. C. de Lamare, "Line-of-sight extra-large mimo systems with angular-domain processing: Channel representation and transceiver architecture," *IEEE Transactions on Communications*, vol. 72, no. 1, pp. 570–584, 2024.
- [15] A. Pizzo, T. L. Marzetta, and L. Sanguinetti, "Spatially-stationary model for holographic mimo small-scale fading," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 1964–1979, 2020.