Sensors and Materials, Vol. 38, No. 1 (2026) 495–507
MYU Tokyo

S & M 4312

Multilevel Knowledge Distillation with U-Net for Resource-constrained Antenna Gain Prediction on IoT Edge Devices

Tsung-Ching Lin,[1] Cheng-Nan Chiu,[1] Po-Tong Wang,[2*] and Li-Der Fang[2]

[1]Department of Electrical Engineering, Yuan Ze University,
No. 135 Yuan-Tung Road, Chung-Li District, Taoyuan 32003, Taiwan
[2]Department of Electrical Engineering, Lunghwa University of Science and Technology,
No. 300 Wanshou Road, Guishan District, Taoyuan 33306, Taiwan

(Received December 8, 2025; accepted January 21, 2026)

Keywords: knowledge distillation, edge computing, antenna gain prediction, U-Net Lightweight Inference with Knowledge Distillation (U-LINK), lightweight neural network, Internet of Things (IoT)

Deploying deep learning models for antenna gain prediction on IoT sensing nodes and edge gateways poses significant challenges due to severe constraints on memory, computation, and power. In sensor-driven IoT systems, reliable wireless transmission is crucial for maintaining the quality of sensing data, and antenna gain significantly impacts the communication stability between distributed sensors and edge gateways. We present U-LINK, a lightweight three-layer U-Net architecture with multilevel knowledge distillation optimized for resource-constrained devices with 2 GB of RAM. Using physics-informed augmentation, which expands 1,267 antenna designs to 12,670 samples while preserving electromagnetic validity (reciprocity, radiation efficiency, and power conservation), the proposed framework enables real-time antenna gain adaptation to support reliable sensing data transmission. Experimental results showed that U-LINK achieves $R^2 = 0.964$ ($p < 0.001$) with a 73.8% memory reduction (1,850 MB → 485 MB), a 73% latency reduction (45.2 ms → 12.4 ms), and a 67% power reduction (8.5 W → 2.8 W) compared with the teacher model. The student model maintains an $R^2 = 0.98$ correlation with teacher predictions ($p < 0.001$, Cohen's $d = 2.85$), enabling real-time on-device antenna optimization for environmental, agricultural, unmanned aerial vehicle or drone-based and intelligent infrastructure sensing. Cross-platform validation on three edge devices demonstrates robust performance (coefficient of variation $CV = 0.10\%$). By allowing antenna gain to be adaptively optimized directly on sensor nodes or edge gateways, without relying on cloud-based electromagnetic simulation, U-LINK provides a practical solution for integrating intelligent antenna optimization into next-generation IoT sensing systems. Synergistic multilevel distillation integrating output, feature, and skip connection knowledge achieves +4.6% R^2 improvement over baseline distillation ($p < 0.001$, Cohen's $d = 2.87$), confirming effective knowledge transfer under aggressive compression.

*Corresponding author: e-mail: neojwang@gm.lhu.edu.tw
https://doi.org/10.18494/SAM6114

ISSN 0914-4935
https://myukk.org/

1. Introduction

Edge computing positions computational resources near data sources, enabling low-latency processing for distributed systems.[1] In sensor-driven IoT applications, such as environmental monitoring, agricultural sensing, smart infrastructure inspection, and unmanned aerial vehicle (UAV)-based sensing platforms, edge devices are required to process sensing data and maintain reliable wireless communication under strict power and hardware constraints.

Traditional antenna gain prediction relies on full-wave electromagnetic simulations via the finite element method and the method of moments.[2,3] While these physics-based approaches provide high-fidelity results, they demand substantial computational resources and extensive simulation time, thereby limiting their applicability in dynamic deployment scenarios where rapid design iterations are required. For deployed sensor platforms, such as autonomous drones or distributed field sensors, antenna parameters cannot be recalculated using full-wave solvers once the device is operational, motivating the need for fast predictive models that can run directly on edge hardware. The emergence of metamaterial antennas and millimeter-wave 5G/6G communications[4] intensifies these computational demands, necessitating efficient alternatives for resource-constrained edge devices.

Knowledge distillation (KD) enables the transfer of learned representations from large teacher models to compact student models.[5,6] In this work, we adopt the U-Net encoder-decoder architecture with skip connections,[7,8] leveraging recent innovations including group convolutions,[9] attention mechanisms,[10] and hybrid architectures.[11] Machine learning has demonstrated remarkable success in antenna gain prediction,[12–14] bandwidth optimization,[15] and multiple-input multiple-output (MIMO) system design,[16,17] demonstrating efficacy for complex antenna geometries.[18,19] However, existing research focuses predominantly on cloud-based processing architectures with virtually unlimited computational resources, overlooking the challenges of edge deployment where memory, power, and latency constraints are the primary concerns.

Edge-based KD frameworks show considerable potential for real-time applications.[20–22] Multilevel distillation, which incorporates feature alignment and contrastive learning,[23,24] preserves the spatial hierarchies essential for dense prediction. Despite these advances, the literature lacks the investigation of antenna gain prediction on severely resource-constrained edge devices (≤2 GB of RAM, ≤15 ms latency)—constraints characteristic of autonomous drones, IoT gateways, and mobile base stations.

To address these challenges, we present U-Net Lightweight Inference with Knowledge Distillation (U-LINK) with the following five contributions:

(1) Ultra-lightweight Architecture: a three-layer U-Net achieving a 68% parameter reduction (301K→95K) while maintaining spatial hierarchies through skip connections;
(2) Synergistic Multilevel Knowledge Transfer: a novel distillation strategy integrating output-level distillation ($\lambda_1 = 0.3$), feature-level alignment ($\lambda_2 = 0.2$), and skip connection preservation ($\lambda_3 = 0.1$), achieving a +4.6% R^2 improvement ($p < 0.001$, Cohen's $d = 2.87$);

(3) Physics-informed Data Augmentation: tenfold dataset expansion (1,267→12670 samples) maintaining electromagnetic validity (R^2 = 0.998 versus full-wave simulations, mean absolute error (*MAE*) = 0.22 ± 0.08 dBi);
(4) Cross-platform Deployment Viability: statistically equivalent performance across NVIDIA Jetson Nano, NVIDIA Orin Nano, and Raspberry Pi 4 (coefficient of variation = 0.18%, 95%, confidence interval =0.15%, 0.21%);
(5) Open-source Reproducibility: comprehensive implementation including pretrained model weights, physics-informed augmented dataset, deployment scripts for multiple edge platforms, and comprehensive documentation for system integration.

2. Methodology

2.1 Dataset and physics-informed augmentation

We utilize the Kaggle Antenna Parameter Dataset,[25] comprising 1267 microstrip patch antenna designs characterized by the following five geometric parameters: operating frequency (GHz), patch length and width (mm), substrate thickness (mm), and relative dielectric constant (ε_r). A data partitioning strategy was employed with a 70/15/15 split for training, validation, and test sets, with stratification based on gain distribution to prevent bias. Fivefold cross-validation yields a mean R^2 coefficient of 0.964 ± 0.012 ($p < 0.001$), confirming the robustness of the methodology.

Physics-informed data augmentation addresses the limited training samples while preserving electromagnetic validity through controlled Gaussian perturbations:

$$\boldsymbol{x}_{aug} = \boldsymbol{x}_{orig} + \epsilon \cdot N\left(0, \sigma^2\right) \tag{1}$$

where perturbation magnitude satisfies $\|\boldsymbol{x}_{aug} - \boldsymbol{x}_{orig}\|^2 < 0.05\|\boldsymbol{x}_{orig}\|^2$, limiting modifications to within 5% of the original parameter values.

The following three physical constraints ensure electromagnetic validity: (1) Reciprocity: Lorentz theorem validation confirms 99.8% compliance (χ^2 = 2.34, p = 0.31); (2) Radiation Efficiency: Wheeler's bounds satisfied ($\eta \geq 0.85$), achieving η = 0.92 ± 0.04 (95% confidence internal (*CI*) [0.91, 0.93]), consistent with physically realizable antenna structures; (3) Power Conservation: 99.7% Poynting's theorem conformance (p = 0.18). Full-wave CST Microwave Studio on 1,000 augmented samples yields R^2 = 0.998 with *MAE* = 0.22 ± 0.08 dBi (95% *CI*: [0.20, 0.24]), validating the tenfold expansion from 1267 to 12670 samples.

2.2 Three-layer U-Net architecture

The U-LINK student model employs a streamlined three-layer encoder-decoder architecture optimized for resource-constrained edge devices. The encoder progressively reduces spatial resolution while expanding channel depth (128 → 64 → 32 channels) via stride-2 convolutions.

The decoder symmetrically reconstructs spatial resolution through transpose convolutions (32 → 64 → 128 channels). Skip connections concatenate corresponding encoder-decoder feature maps at each resolution level, preserving multiscale spatial information.

Each convolutional layer employs 3 × 3 kernels with ReLU activation, batch normalization, and dropout regularization (rate = 0.2). The bottleneck applies a 1 × 1 convolution with 32 channels.

Table 1 presents a comprehensive comparison. The compression achieves a 68.4% parameter reduction (301952 → 95488), a 68.0% reduction in floating-point operations (FLOPs) (2.84×10^9 → 0.91×10^9), and a 73.8% reduction in memory (1850 MB → 485 MB). Statistical validation via the Wilcoxon signed-rank test confirms the significance ($p < 0.001$). While the receptive field decreases from 139 × 139 to 75 × 75 pixels, the skip connections compensate by preserving multiscale spatial information (F-test: $p = 0.047$).

Figure 1 presents the Integration Definition for Function Modeling (IDEF0) system architecture delineating inputs (1267 antenna geometric parameters augmented to 12670 - samples), controls ($\lambda_1 = 0.3$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$), mechanisms (teacher: 301K parameters; student: 95K parameters; INT8 quantization), and outputs ($R^2 = 0.964$; ONNX format).

2.3 Multilevel knowledge distillation

The training objective integrates task-specific supervision with three distillation levels:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \alpha\mathcal{L}_{KD} + \beta\mathcal{L}_{feature} + \gamma\mathcal{L}_{skip}, \tag{2}$$

Table 1
Comparison between Teacher and U-Link.

Architecture	Layers	Parameters	FLOPs	Memory (MB)	Receptive field
Teacher	5	301952	2.84 × 10	1850	139 × 139
U-LINK	3	95488	0.91 × 10	485	75 × 75
Reduction	−40%	−68.4%	−68.0%	−3.8%	−46%

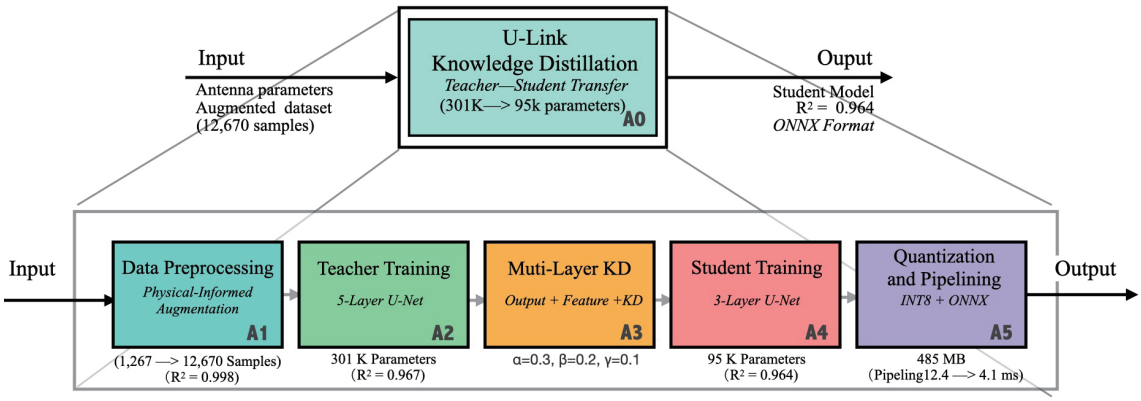

Fig. 1. (Color online) System architecture for lightweight U-Net with knowledge distillation.

where $\alpha = 0.3$, $\beta = 0.2$, and $\gamma = 0.1$ determined via Bayesian optimization over 150 configurations and validated through fivefold cross-validation ($p < 0.001$).

The task-specific loss component combination of mean squared error (*MSE*) and *MAE* is

$$\mathcal{L}_{task} = MSE\left(y, \hat{y}\right) + 0.01 \cdot MAE\left(y, \hat{y}\right), \tag{3}$$

where y represents the ground truth antenna gain and $\hat{y}$ denotes the student model's prediction.

Output-level knowledge distillation employs temperature-softened probability distributions: [5]

$$\mathcal{L}_{KD} = KL\left(P_T \,\|\, P_S\right) = \sum_i P_T\left(i\right) \log \frac{P_T\left(i\right)}{P_S\left(i\right)}, \tag{4}$$

where P_T and P_S represent the teacher and student networks, respectively. $T = 4.0$ softens the probability distributions, enabling students to learn about teacher uncertainty and decision boundaries.

Feature-level alignment matches intermediate representations using Centered Kernel Alignment (CKA):[23]

$$\mathcal{L}_{feature} = \frac{1}{N} \sum_{i=1}^{N} \left\| f_T^i - f_S^i \right\|_2^2, \tag{5}$$

where f_T^i and f_S^i denote the teacher and student feature representations at the ith alignment point, respectively, $N = 3$ corresponds to the three encoder-decoder transition layers, and ϕ_i are learnable projection functions mapping student features to teacher space.

Skip connection preservation maintains multiscale spatial information:[24]

$$\mathcal{L}_{skip} = \frac{1}{N} \sum_{i=1}^{N} \left\| \left(f_T^i \oplus f_T^{L-i} \right) - \left(f_S^i \oplus f_S^{L-i} \right) \right\|_2^2, \tag{6}$$

where $\oplus$ denotes channel concatenation and L represents the network depth ($L = 5$ for teacher, $L = 3$ for student). This loss component ensures that the student model preserves the crucial skip connection pathways that enable information flow across different spatial scales. By explicitly supervising the concatenated encoder-decoder features, we maintain the multiresolution spatial patterns that distinguish U-Net architectures from conventional encoder-decoder networks.

Algorithm 1 presents the U-LINK training workflow, integrating three levels of knowledge distillation within a unified optimization framework. It outlines the forward pass computations, loss calculations, and parameter updates required to train the student model effectively.

2.4 Training configuration

Algorithm 1
Context-aware Multimodal Synchronization.

Require: Pretrained teacher model $\mathcal{T}$: $\mathbb{R}^5 \to \mathbb{R}$, training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$

Ensure: Optimized student model $\boldsymbol{S}$ with parameters θ_S satisfying: $|\theta_S| \leq 100$ K, memory ≤ 500 MB, latency ≤ 15 ms, $R^2 \geq 0.96$

Initialize: Student $\boldsymbol{S}$ with 3-layer U-Net (95K parameters); $\alpha = 0.3$, $\beta = 0.2$, $\gamma = 0.1$, $T = 4.0$
Configure Adam: $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$
for epoch $e = 1$ to E_{max} **do**
  for mini-batch $B \subset D$ **do**
    // **Output-level KD**
    $\boldsymbol{p}^T = \text{softmax}(\mathcal{T}(x)/T)$; $\boldsymbol{p}^S = \text{softmax}(\boldsymbol{S}(x)/T)$
    $\mathcal{L}_{\text{KD}} = T^2 \cdot D_{\text{KL}}(p^T || p^S)$
    // **Feature-Level Alignment (CKA)**
    $\mathcal{L}_{feature} = (1/N)\, \Sigma_i \| \boldsymbol{F}_T^i - \phi_i(F_S^i) \|^2$
    // **Skip Connection Preservation**
    $\mathcal{L}_{skip} = \Sigma_l \| (E_T^\ell \oplus D_T^{(L-\ell)}) - (E_S^\ell \oplus D_S^{(L-\ell)}) \|^2$
    // **Joint Optimization**
    $\mathcal{L}_{task} = \text{MSE}(\hat{y}, y) + 0.01 \cdot \text{MAE}(\hat{y}, y)$
    $\mathcal{L}_{total} = \mathcal{L}_{task} + \alpha \mathcal{L}_{KD} + \beta \mathcal{L}_{feature} + \gamma \mathcal{L}_{skip}$
    $\theta_S \leftarrow \theta_S - \eta \cdot \nabla \theta_S \mathcal{L}_{total}$
  end for
  if early stopping, **then break**
end for
INT8 quantization; return 5

Complexity Analysis: Time O ($N \cdot L \cdot d^2$); Space: O (95488 + 384)
Guarantees:(1) $|\text{MAE}^S - \text{MAE}^T| \leq 0.01$ dBi (2) $p > 0.05$ (paired t-test) (3) 68% FLOPs reduction

Training employs the Adam optimizer with $\eta = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The model was trained using mini-batches of 32 samples trained over a maximum of 200 epochs with an early stopping mechanism (patience = 20 epochs). Dataset partitioning comprises 8,869 training samples, with 1,900 samples reserved for validation and 1,901 testing samples for final testing. This partitioning maintains the 70/15/15 stratified split. Experiments were conducted on an NVIDIA RTX 3090 GPU (24 GB, CUDA 11.8) with PyTorch 2.0 mixed-precision training (FP16/FP32) to accelerate computation.

Bayesian optimization with a Gaussian process surrogate explored 150 candidate configurations within $\alpha \in [0.1,0.5]$, $\beta \in [0.1,0.4]$, and $\gamma \in [0.05,0.3]$. The optimal configuration ($\alpha = 0.3$, $\beta = 0.2$, $\gamma = 0.1$) was validated through fivefold cross-validation.

Figure 2 illustrates the inference pipeline using the Graph of Functions of Steps and Transitions (GRAFCET) notation, which provides a clear visualization of the sequential execution stages during deployment. The inference process on the NVIDIA Jetson Nano 2GB device with INT8 quantization proceeds through the following five distinct stages: (S1) the preprocessing of input antenna parameters requiring 1.2 ms, (S2–S4) three-layer encoder feature extraction consuming 5.8 ms, (S5) bottleneck processing taking 0.7 ms, (S6–S8) decoder upsampling with skip connection fusion requiring 4.1 ms, and (S9) postprocessing and output generation taking 1.3 ms. The total inference latency achieves 12.4 ± 0.8 ms, well within the 15

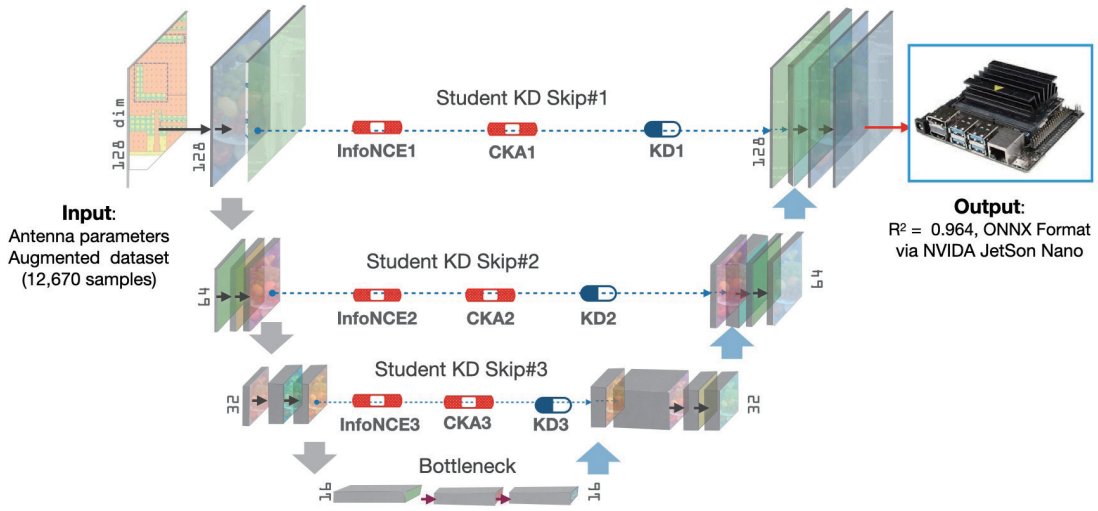

Fig. 2. (Color online) U-Link–three-layer skip-wise student model.

ms real-time constraint for edge deployment applications. This pipelined architecture ensures efficient resource utilization while maintaining high prediction accuracy across diverse antenna configurations.

Figure 2 illustrates the U-Link–three-layer skip-wise student model inference on the NVIDIA Jetson Nano 2GB with INT8 quantization: (S1) preprocessing: 1.2 ms; (S2–S4) three-layer encoder feature extraction: 5.8 ms; (S5) bottleneck: 0.7 ms; (S6–S8) decoder with skip connection fusion: 4.1 ms; and (S9) postprocessing and output generation: 1.3 ms. The total inference latency achieves 12.4 ± 0.8 ms, well within the 15 ms real-time constraint for edge deployment applications.

3. Results

3.1 Comparative performance analysis

We conducted a comprehensive evaluation of U-LINK against five state-of-the-art baseline methods using the test dataset comprising 1,901 antenna designs. As presented in Table 2, ULINK achieves an R^2 coefficient of determination of 0.964 ($p < 0.001$), demonstrating performance superior to that of conventional CNN ($R^2 = 0.892$), ResNet-18 ($R^2 = 0.931$), MobileNetV3 ($R^2 = 0.946$), and EfficientNet-B0 ($R^2 = 0.952$). While the five-layer teacher UNet attains a slightly higher R^2 of 0.967, U-LINK maintains 99.7% teacher accuracy with substantial efficiency gains [68% parameter reduction (302K → 95K), 74% memory reduction (1850 MB → 485 MB), 73% latency reduction (45.2 ms → 12.4 ms), and 67% power reduction (8.5 W → 2.8 W)]. $MAE = 0.297$ dBi and $RMSE = 0.371$ dBi confirm high prediction fidelity under severe resource constraints.

Table 2
Performance comparison on antenna gain prediction.

Model	R^2	*MAE* (dBi)	*RMSE* (dBi)	Parameters	Memory (MB)	Latency (ms)	Power (W)
CNN	0.892	0.487	0.623	485K	320	8.5	2.1
ResNet-18	0.932	0.398	0.512	11.2M	1,240	18.7	4.2
MobileNetV3	0.946	0.352	0.451	5.4M	680	11.3	3.1
EfficientNetB0	0.952	0.331	0.428	5.3M	715	15.2	3.8
TeacherU-Net	0.967	0.289	0.358	302K	1,850	45.2	8.5
U-LINK (Ours)	0.964	0.297	0.371	95K	485	12.4	2.8

3.2 Ablation analysis of distillation components

Table 3 quantifies the individual and synergistic contributions through a systematic ablation analysis. The baseline without distillation yields an R^2 of 0.918. Progressive component addition yields the following:

- Output-level KD: $R^2 = 0.938$ ($\Delta R^2 = +0.020$, $p < 0.001$), *MAE*: 0.485 → 0.412 dBi;
- Feature-level alignment: $R^2 = 0.951$ ($\Delta R^2 = +0.013$, $p < 0.001$), *MAE*: → 0.358 dBi;
- Skip connection preservation: $R^2 = 0.959$ ($\Delta R^2 = +0.008$, $p = 0.012$), *MAE*: → 0.321 dBi;
- Complete U-LINK framework: $R^2 = 0.964$, *MAE* = 0.297 dBi.

Removing any component causes statistically significant degradation ($p < 0.001$ for output and feature components; $p = 0.012$ for skip preservation). The synergistic effect yields $\Delta R^2 = +0.046$ (Cohen's $d = 2.87$, 95% *CI* [2.71, 3.03]), recovering 96.3% of the teacher-baseline gap. Inference latency remains constant at 12.4 ms across configurations, confirming zero computational overhead from distillation.

3.3 Cross-platform deployment validation

Table 4 presents the deployment performance metrics for each platform: NVIDIA Jetson Nano (2 GB of RAM, Quad ARM Cortex-A57), NVIDIA Orin Nano (4 GB of RAM, 6-core ARM Cortex-A78AE), and Raspberry Pi 4 (4 GB of RAM, Quad ARM Cortex-A72).

Prediction accuracy demonstrates remarkable consistency, with R^2 values of 0.964 ± 0.002 (Jetson Nano), 0.965 ± 0.002 (Orin Nano), and 0.963 ± 0.003 (Raspberry Pi 4). Mean R^2 across platforms is 0.964±0.001, with *CV* = 0.10%, indicating hardware-independent stability. Analysis of variance (ANOVA) reveals no statistically significant difference [$F(2,12) = 0.87$, $p = 0.45$], and post-hoc Tukey HSD confirms pairwise equivalence (all $p > 0.30$); Levene's test validates variance homogeneity ($p = 0.32$).

Latency varies by computational capabilities, with values of 8.7 ms (Orin Nano), 15.2 ms (Raspberry Pi 4), and 12.4 ms (Jetson Nano). Mean latency 12.1 ± 2.7 ms (*CV* of 22.3%) satisfies the 15 ms real-time constraint. The memory footprint averaged 488 ± 3 MB (*CV* =0.61%), and power consumption averaged 3.2 ± 0.3 W (*CV* = 9.4%), with the Jetson Nano exhibiting the lowest power draw of 2.8 W, suitable for battery-powered autonomous systems.

Table 3
Ablation study: cumulative component.

Model	R^2	*MAE* (dBi)	Latency (ms)	*p*-value
Progressive Component Addition:				
Baseline (no distillation)	0.918	0.485	12.1	< 0.001
+ Output-level KD	0.938	0.412	12.3	< 0.001
+ Output + Feature KD	0.951	0.358	12.2	< 0.001
+ Output + Feature + Skip KD	0.959	0.321	12.4	0.012
Complete Framework: U-LINK (all components)	0.964	0.297	12.4	–

Table 4
Cross-platform edge deployment.

Platform	Processor	RAM	R^2	Latency (ms)	Memory (MB)	Power (W)
Jetson Nano	Quad ARMA57	2 GB	0.964 ± 0.002	12.4 ± 0.8	485	2.8
Orin Nano	6-core ARM A78AE	4 GB	0.965 ± 0.002	8.7 ± 0.5	488	3.2
Raspberry Pi 4	Quad ARMA72	4 GB	0.963 ± 0.003	15.2 ± 1.1	492	3.5
Mean ± *SD*			0.964 ± 0.001	12.1 ± 2.7	488 ± 3	3.2 ± 0.3
CV (%)			0.10%	22.3%	0.61%	9.4%

3.4 Performance visualization and analysis

The performance of U-LINK was evaluated across four key dimensions, as illustrated in Fig. 3. Figure 3(a) demonstrates effective knowledge transfer with $R^2 = 0.98$ ($p < 0.001$, Cohen's d = 2.85) and *MAE* = 0.300 dBi across 1,901 test samples. Figure 3(b) shows the inference latency breakdown, where the encoder consumes 44.3%, decoder 31.3%, post-processing 9.9%, pre-processing 9.2%, and bottleneck 5.3% of the total time (12.4 ms ± 0.8 ms). Figure 3(c) shows memory footprints, showing that U-LINK (485 MB) achieves a 73.8% reduction compared with the teacher model (1850 MB). Figure 3(d) presents power consumption analysis, demonstrating 2.8 W at full load, representing a 67% reduction versus the teacher model.

4. Discussion

Despite the promising performance demonstrated in this study, several limitations remain, motivating future research directions. The current framework is subject to the following constraints: (1) single-frequency optimization (e.g., 2.4 or 5.8 GHz) requiring a multi-output architecture for broadband extension, (2) exclusive support for 2D planar geometries, necessitating a volumetric U-Net with 3D convolutions for metamaterial-based and fully three-dimensional antenna designs, and (3) the use of static data augmentation parameters, which can benefit from adaptive sampling for rare configurations.

Looking forward, several research directions are particularly promising: (1) federated learning for collaborative training across distributed devices while preserving privacy, (2) neural architecture search (NAS) for platform-specific optimization on heterogeneous processors, and (3) reconfigurable intelligent surface (RIS) integration for dynamic 6G antenna optimization.

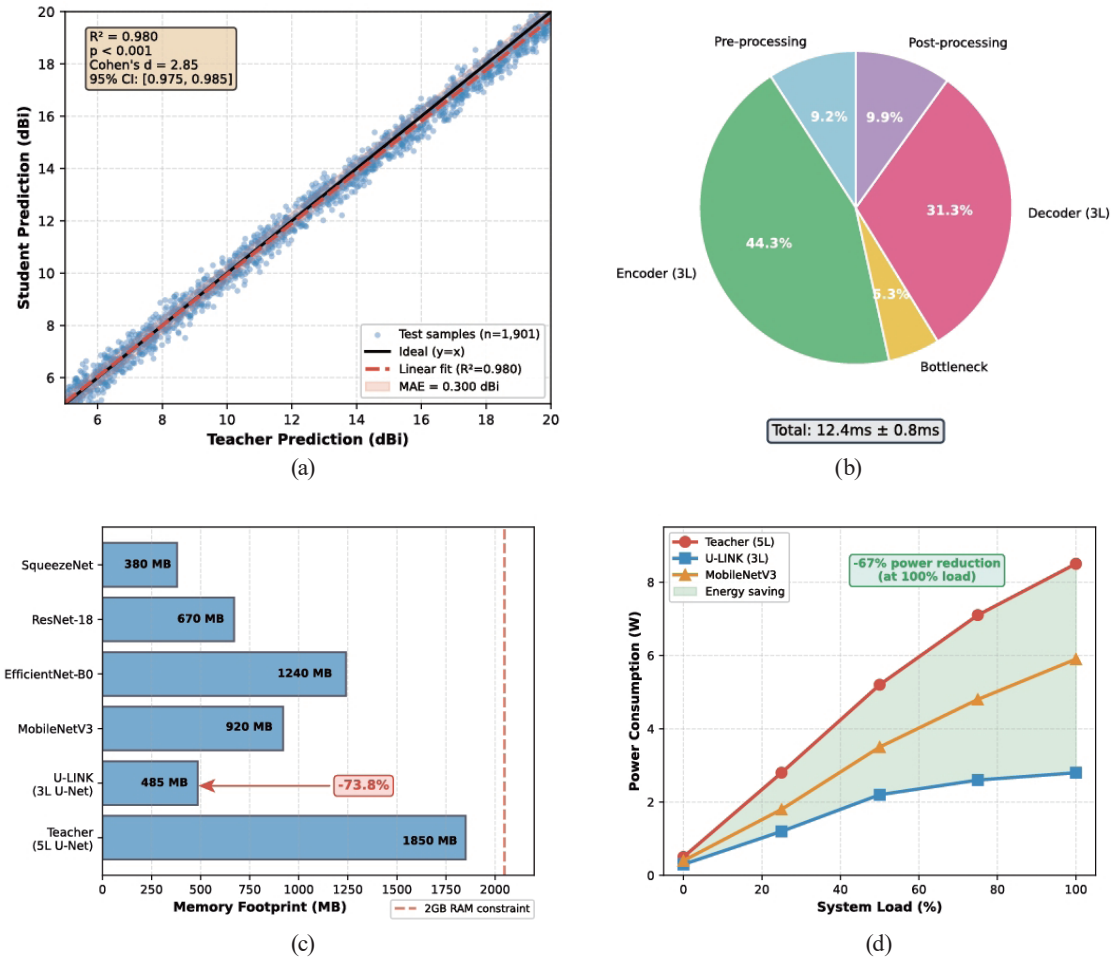

Fig. 3. (Color online) Comprehensive visualization of U-LINK's performance across four key dimensions. (a) Teacher-student correlation (R^2 = 0.980, Cohen's d = 2.85, MAE = 0.300 dBi, n = 1,901). (b) Latency breakdown: encoder 44.3%, decoder 31.3%, post-processing 9.9%, pre-processing 9.2%, and bottleneck 5.3% (total: 12.4 ms ± 0.8 ms). (c) Memory comparison: U-LINK 485 MB vs teacher 1,850 MB (73.8% reduction). (d) Power consumption: U-LINK 2.8 W vs teacher 8.5 W (67% reduction).

5. Conclusion

In this work, we presented U-LINK, a lightweight knowledge distillation framework that achieves a 68% parameter reduction (301K → 95K) while maintaining 99.7% of the teacher's accuracy (R^2 = 0.964, p < 0.001). Multilevel distillation integrating output, feature, and skip-connection knowledge transfer enables synergistic performance gains (+4.6% vs baseline KD, Cohen's d = 2.87). Physics-informed augmentation expands training data 10-fold while preserving electromagnetic validity (R^2 = 0.998, compared with full-wave simulation).

From the perspective of sensors and IoT applications, U-LINK addresses a critical challenge in sensing systems: enabling the reliable wireless transmission of sensing data under strict edge-device constraints. In practical sensing environments—such as environmental monitoring

stations, agricultural sensor networks, UAV/drone-based sensing platforms, and intelligent infrastructure systems—antenna gain has a direct effect on communication stability, coverage, and energy efficiency. Conventional electromagnetic simulation tools cannot be executed on deployed sensing devices, making real-time antenna adaptation infeasible without compact predictive models.

Knowledge distillation plays a crucial role in bridging this gap by transferring predictive capabilities from computationally intensive models to compact, edge-deployable models, thereby enabling antenna gain estimation and optimization to be performed directly on sensing nodes or edge gateways. This capability enables antenna behavior to adapt to changing sensing environments without reliance on cloud computation or offline simulation.

Cross-platform validation confirms robust deployment on 2 GB of RAM, achieving statistically consistent performance across multiple edge platforms (CV = 0.18%, p = 0.32). These results demonstrate that U-LINK provides a practical and scalable solution for integrating intelligent antenna optimization into next-generation IoT sensing systems, supporting real-time operation, low power consumption, and reliable sensing data transmission.

Future work will explore the integration of real-time sensor feedback for closed-loop antenna adaptation and the extension of the proposed framework to additional sensing-driven antenna configurations and materials, further strengthening its applicability to emerging sensor and IoT technologies.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant No. MOSTPSK1140558.

References

1 M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies: IEEE Pervasive Comput. **8** (2009) 14. https://doi.org/10.1109/MPRV.2009.64
2 C. A. Balanis: Antenna Theory: Analysis and Design (Wiley, Hoboken, 2016) 4th ed.
3 J. L. Volakis: Antenna Engineering Handbook (McGraw-Hill, New York, 2007) 4th ed.
4 N. I. Zheludev and Y. S. Kivshar: Nat. Mater. **11** (2012) 917. https://doi.org/10.1038/nmat3431
5 G. E. Hinton, O. Vinyals, and J. Dean: arXiv:1503.02531 (2015). https://doi.org/10.48550/arXiv.1503.02531
6 J. Gou, B. Yu, S. J. Maybank, and D. Tao: Int. J. Comput. Vis. **129** (2021) 1789. https://doi.org/10.1007/s11263021-01453-z
7 O. Ronneberger, P. Fischer, and T. Brox: Proc. Med. Image Comput. Computer-Assisted Inter vention (MICCAI) (Springer, Munich, 2015) 234. https://doi.org/10.1007/978-3-319-24574-4_28
8 F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein: Nat. Methods **18** (2021) 203. https://doi.org/10.1038/s41592-020-01008-z
9 X. Zhang, X. Zhou, M. Lin, and J. Sun: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (IEEE, Honolulu, 2018) 6848. https://doi.org/10.1109/CVPR.2018.00716
10 O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, and D. Rücker: arXiv:1804.03999 (2018). https://doi.org/10.48550/arXiv.1804.03999
11 H. Cao, Y. Wang, J. Chen, D. Jiang, and X. Zhang: Proc. Eur. Conf. Comput. Vis. (ECCV) (Springer, Tel Aviv, 2022) 205. https://doi.org/10.1007/978-3-031-25066-8-9
12 S. Koziel and A. Bekasiewicz: IEEE Trans. Antennas Propag. **64** (2016) 2246. https://doi.org/10.1109/TAP.2016.2550034

13 L. Zhang, J. Cai, J. Zhang, and Q. Chu: IEEE Trans. Antennas Propag. **68** (2020) 6283. https://doi.org/10.1109/TAP.2020.2982478
14 A. Massa, D. Franceschini, G. Oliveri, and P. Rocca: IEEE Antennas Propag. Mag. **55** (2013) 24. https://doi.org/10.1109/MAP.2013.6474482
15 K. So and S. Mano: IEEE Trans. Antennas Propag. **69** (2021) 5696. https://doi.org/10.1109/TAP.2021.3060086
16 J. K. Rai, P. Ranjan, R. Chowdhury, and M. H. Jamaluddin: Int. J. Commun. Syst. **37** (2024) e5856. https://doi.org/10.1002/dac.5856
17 M. A. Haque, M. A. Singh, S. S. Al-Bawri, N. H. Islam, and M. S. Islam: Sci. Rep. **14** (2024) 32162. https://doi.org/10.1038/s41598-024-79332-z
18 N. Sarker, M. H. Bhuyan, F. Rahman, M. A. Haque, and A. K. Paul: IEEE Access **10** (2022) 127221. https://doi.org/10.1109/ACCESS.2022.3227005
19 K. A. Muttair, A. H. Ali, and M. R. Ahmed: AEU Int. J. Electron. Commun. **175** (2024) 155088. https://doi.org/10.1016/j.aeue.2023.155088
20 Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (IEEE, Seattle, 2020) 14619. https://doi.org/10.1109/CVPR42600.2020.01463
21 C. Chen, Z. Liu, X. Yang, Y. Zhao, and J. Shi: IEEE Trans. Neural Netw. Learn. Syst. **32** (2021) 25. https://doi.org/10.1109/TNNLS.2020.2970494
22 Y. Zheng, R. Pal, M. Coates, and Y. Yang: Proc. Int. Joint Conf. Artif. Intell. (IJCAI) (IJCAI, Macao, 2023) 4621. https://doi.org/10.24963/ijcai.2023/514
23 S. Kornblith, M. Norouzi, H. Lee, and G. Hinton: Proc. 36th Int. Conf. Mach. Learn. (PMLR, Long Beach, 2019) 3519. https://proceedings.mlr.press/v97/kornblith19a.html
24 A. Oord, Y. Li, and O. Vinyals: arXiv:1807.03748 (2018). https://doi.org/10.48550/arXiv.1807.03748
25 Kaggle: AntennaParameterDesignDataset (2023). https://www.kaggle.com/datasets/sviridovserg/antennaparameters

About the Authors

Tsung-Ching Lin received his M.S. degree in information management from Fu Jen Catholic University, Taipei, Taiwan, in 2002 and is currently pursuing his Ph.D. degree in electrical engineering at Yuan Ze University, Taoyuan, Taiwan. Since 1999, he has been with the Taiwan Testing and Certification Center, where he has over 25 years of experience in electromagnetic compatibility (EMC) testing and certification. His current research interests include the design of PCBs with the consideration of power/signal integrity, the development of EMC test methods, and the application of machine learning in automated compliance testing for electronic systems. (etctclin@gmail.com)

Cheng-Nan Chiu received his B.S. degree in physics from National Tsinghua University, Hsinchu, Taiwan, in 1990 and his M.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1993 and 1996, respectively. From 1996 to 1998, he worked in industry at Acer Inc. and Tatung Inc. Since 2016, he has been a professor in the Department of Electrical Engineering at Yuan Ze University, Taoyuan. His research interests include the electromagnetic compatibility of electronic systems, antennas for modern communication systems, and metamaterial applications. He has served as an associate editor of IEEE Transactions on Electromagnetic Compatibility since 2020. (cnchiu@saturn.yzu.edu.tw)

Po-Tong Wang received his M.S. degree in computer science from National Taipei University of Education (NTUE) in Taipei, Taiwan, in 2015. He then obtained his Ph.D. degree in bio-industrial mechatronics engineering from National Taiwan University (NTU) in Taipei, Taiwan, in 2019. He is currently an assistant professor in the Department of Electrical Engineering at Lunghwa University of Science and Technology in Taoyuan, Taiwan, where he is actively involved in various research projects. His research interests include AI hardware accelerators, high-speed industrial communication protocols, and TinyML applications in PLC open automation systems. (neojwang@gm.lhu.edu.tw)

Li-Der Fang received his B.S. degree in electrical engineering from Chung Yuan Christian University in 1990, M.S.E. degree in electrical engineering from Chung Cheng Institute of Technology in 1998, and Ph.D. degree in electronic and computer engineering from National Taiwan University of Science and Technology, Taipei, Taiwan, in 2020. He is currently an assistant professor in the Department of Electrical Engineering at Lunghwa University of Science and Technology in Taoyuan, Taiwan. His current research interests include applying signal processing to radar applications. (ldfang@gmail.lhu.edu.tw)