Coverage Prediction for mmWave Mobile Communication Systems at the Edge: A Vision Transformer Approach

Leran Qi and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—Coverage prediction is the pivotal quality-of-service indicator in cellular systems, which typically relies on propagation models for path loss estimation, leading to substantial computational overhead. In this letter, we present an efficient vision transformer (ViT)-based framework to directly capture correlations among terrain, interference, and coverage characteristics, bypassing the conventional path loss modeling. Experiments on urban scenarios demonstrate that ViT achieves the lowest prediction error against the ray tracing benchmark, outperforming COST 231 by more than 10% and exhibiting lower error than convolutional neural network (CNN)-based models, including AlexNet, ResNet, and a recent CNN for propagation prediction. Moreover, it also reduces computation time by up to two orders of magnitude compared to propagation models. This highlights the potential of our proposed method for low-latency coverage prediction with minimal prediction error in resource-constrained edge environments.**

***Index Terms*—Coverage prediction, edge computing, network optimization, vision transformer.**

I. Introduction

With the accelerated evolution of 5G and the advent of 6G, mobile communication networks are under mounting pressure to support massive device connectivity, deliver ultra-low latency, and adapt to diverse and complex deployment scenarios. To address spectrum scarcity and support escalating capacity demands, these networks are progressively adopting higher frequency bands. In particular, millimeter-wave (mmWave) communication has gained prominence as a key enabler, offering large bandwidth, high data rates, and support for spatial multiplexing. Nevertheless, severe path loss and acute blockage sensitivity in mmWave signals necessitate denser base station deployments to ensure service quality [1]. Moreover, the highly variable propagation characteristics of mmWave signals introduce uncertainty into network behavior. Concurrently, network architectures are shifting toward edge computing, bringing computation closer to end users to improve responsiveness [2]. The propagation uncertainty, network densification, and dynamic edge environments collectively complicate the network operating environment and pose new challenges for network optimization [3], [4].

Wireless network optimization generally follows an iterative process comprising network definition, coverage analysis, and performance assessment. Within this process, coverage prediction plays a pivotal role by quantifying service availability and enabling informed optimization decisions. Specifically, coverage rate is defined as the proportion of locations within a target service area where key signal quality metrics, i.e., the maximum received power and the signal-to-interference ratio (SIR), meet the predefined thresholds. In edge computing scenarios, coverage prediction becomes particularly demanding, as it must deliver both high accuracy and low latency under dynamic conditions and stringent resource constraints.

Traditional methods for coverage prediction are typically built upon propagation models, which are classified into empirical and analytical [5]. Empirical models, such as *Okumura* model [6] and *Hata* model [7], are constructed from extensive field measurements, enabling fast predictions in similar environments, but their accuracy degrades in heterogeneous terrains or dynamic scenarios. Analytical models, such as ray tracing and finite-difference time-domain methods, achieve higher fidelity through detailed electromagnetic calculations, but at the cost of substantial computational overhead, which restricts their suitability for latency-sensitive edge computing [8], [9]. Recently, machine learning-based approaches have been adopted to enhance prediction using historical data [10]. However, their inability to capture long-range spatial dependencies limits their generalization across diverse environments. Moreover, both traditional and learning-based methods focus on predicting signal strength, requiring additional efforts to derive coverage rate, which increases complexity and risks compounding errors.

To balance accuracy and efficiency in coverage prediction, in this letter, we introduce a ViT-based deep learning framework that establishes an end-to-end mapping from terrain information, building layout and antenna features to spatial coverage distributions [11]. Unlike CNNs that focus on local features, the transformer architecture employs self-attention to capture long-range spatial dependencies and complex environmental patterns, which is well suited for coverage prediction since coverage rate reflects a global indicator of signal quality [12]. Tailored to the real-time and resource-constrained demands of edge computing, the proposed method directly predicts coverage metrics without explicit path loss modeling, thereby enabling fast, accurate performance evaluation in edge network optimization. Experiments demonstrate clear advantages over traditional propagation models as well as CNN-based approaches, underscoring both the novelty of our method and its potential for future mobile networks.

Received 15 July 2025; revised 15 September 2025; accepted 24 September 2025. Date of publication 1 October 2025; date of current version 24 February 2026. The associate editor coordinating the review of this article and approving it for publication was C. Cicconetti. *(Corresponding author: Shaowei Wang.)*

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: leran.qi@smail.nju.edu.cn; wangsw@nju.edu.cn).

Digital Object Identifier 10.1109/LNET.2025.3616319

II. System Model

Consider a mobile communication network with each base station equipped with three directional antennas. The base station serves a local area partitioned into square grids, where the received power at each grid center serves as a proxy for the average signal strength across the grid. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the sets of grids and antennas, respectively. For each antenna $n \in \mathcal{N}$, the azimuth $\theta_n \in \Theta$ is the angle between the projection of its main lobe direction onto the horizontal plane and true north; the downtilt $\phi_n \in \Phi$ is the angle between the main lobe and the horizontal plane.

Let P_t denote the antenna transmit power; $G_{n,k}$ the transmission gain from antenna n to grid k, dependent on its azimuth and downtilt; G_k the receiver antenna gain at grid k; and $L_{n,k}$ the path loss from antenna n to grid k. The received reference signal power $P_{k,n}$ at grid k from antenna n can be expressed as

$$P_{k,n} = P_t + G_{n,k} + G_k - L_{n,k}. \tag{1}$$

The maximum reference signal power at grid k from all antennas is defined as $P_k = \max_n P_{k,n}$. Grid k is served by the antenna providing this maximum power, while signals from other antennas are treated as interference.

A grid k is covered effectively if

$$P_k \geq P_{\text{th}}, \tag{2}$$

and

$$\frac{P_k}{\sum_n P_{k,n} - P_k} \geq \text{SIR}_{\text{th}}. \tag{3}$$

Eq. (2), commonly referred to as the power coverage criterion, ensures the received signal strength exceeds the minimum threshold required for successful signal demodulation at the receiver hardware. Eq. (3), known as the capacity coverage constraint, maintains sufficient communication quality by preventing excessive interference-induced degradation. Simply increasing the transmission power of antennas can indeed expand the power coverage area, but it may also lead to interference between the signals from different antennas. Therefore, network optimization must strike a balance between power coverage and capacity coverage. The overall coverage rate is calculated as

$$\frac{\sum_{k\in\mathcal{K}} \mathbb{I}\Big(P_k \geq P_{\text{th}} \text{ and } \frac{P_k}{\sum_n P_{k,n} - P_k} \geq \text{SIR}_{\text{th}}\Big)}{|\mathcal{K}|}, \tag{4}$$

where $|\mathcal{K}|$ denotes the cardinality of set $\mathcal{K}$, and $\mathbb{I}(x)$ is the indicator function defined as $\mathbb{I}(x) = 1$ if condition x is true, and 0 otherwise.

The objective of the coverage prediction problem is to predict the coverage rate calculated by (4), given the distribution of ground buildings in the wireless communication network, the azimuth and downtilt of antenna n_0, and the signal strength received by each grid from other antennas $n' \in \mathcal{N} \setminus \{n_0\}$.

III. Proposed Method

A. Dataset Construction

To construct the learning dataset, a real city map is partitioned into multiple fixed-size local maps, each assumed to contain one base station. For each local map, the received signal strength at each grid location is calculated from all antennas in the network. A naive approach would be to use the signal received from each antenna as a separate input feature. However, this leads to increased input dimensionality as the number of base stations grows, thereby raising model complexity and training cost. Moreover, since coverage rate computation depends only on the aggregated signal profile rather than individual antenna contributions, such detailed input is redundant. To address this, the signal strength received by a grid from antennas $n' \in \mathcal{N}\setminus\{n_0\}$ is compressed into two features: the maximum received signal $\max_{n'\in\mathcal{N}\setminus\{n_0\}} P_{k,n'}$ and the total received power $\sum_{n'\in\mathcal{N}\setminus\{n_0\}} P_{k,n'}$.

The neural network input consists of four feature maps: the building distribution within the local map, the signal strength received from antennas other than the target antenna n_0, and the azimuth and downtilt angles of the target antenna n_0. The model predicts the coverage of the local map, with labels derived from coverage calculated based on received signal powers $P_{k,n}$ at each grid k from antenna n. These received powers are generated by a 3D ray tracing model implemented in the radio propagation software *Winprop* [13]. Following the official documentation, we set the subdivision size of walls and wedges to 20 m, applied an adaptive resolution factor of 2, used a spherical zone radius of 150 m, and enabled multiple reflections and diffractions.

B. Network Architecture

The input data in coverage prediction tasks can be represented as multi-channel images, thus we apply ViT for coverage prediction. Let the input data be denoted as $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, where H and W denote the number of grid cells in the horizontal and vertical directions of the local map, respectively, and $C = 4$ denotes the number of channels. ViT first divides $\boldsymbol{x}$ into patches of size $P \times P$, resulting in $\boldsymbol{x'} \in \mathbb{R}^{(HW/P^2)\times(P^2\cdot C)}$. Each patch is then flattened, linearly projected, and position encoded to obtain the first layer input for the transformer encoder:

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}'_1\boldsymbol{E}; \boldsymbol{x}'_2\boldsymbol{E}; \ldots; \boldsymbol{x}'_N\boldsymbol{E}\right] + \boldsymbol{E}_{\text{pos}}, \tag{5}$$

where $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_N$ are the patches obtained by dividing $\boldsymbol{x}$, $\boldsymbol{E}$ is the linear projection matrix, and $\boldsymbol{E}_{\text{pos}}$ is the positional encoding.

ViT contains L layers of transformer encoders. Each encoder consists of two sub-layers. The first sub-layer applies layer normalization and multi-head self-attention, whose operations are denoted as $f_{\text{LN}}(\cdot)$ and $f_{\text{MSA}}(\cdot)$, respectively. The second sub-layer contains layer normalization followed by a fully connected layer, denoted by the function $f_{\text{MLP}}(\cdot)$. Each sub-layer incorporates a residual connection, formally expressed as

$$\begin{aligned} \boldsymbol{z}'_\ell &= f_{\text{MSA}}(f_{\text{LN}}(\boldsymbol{z}_{\ell-1})) + \boldsymbol{z}_{\ell-1}, \quad \ell = 1, \ldots, L, \\ \boldsymbol{z}_\ell &= f_{\text{MLP}}\big(f_{\text{LN}}\big(\boldsymbol{z}'_\ell\big)\big) + \boldsymbol{z}'_\ell, \quad \ell = 1, \ldots, L. \end{aligned} \tag{6}$$

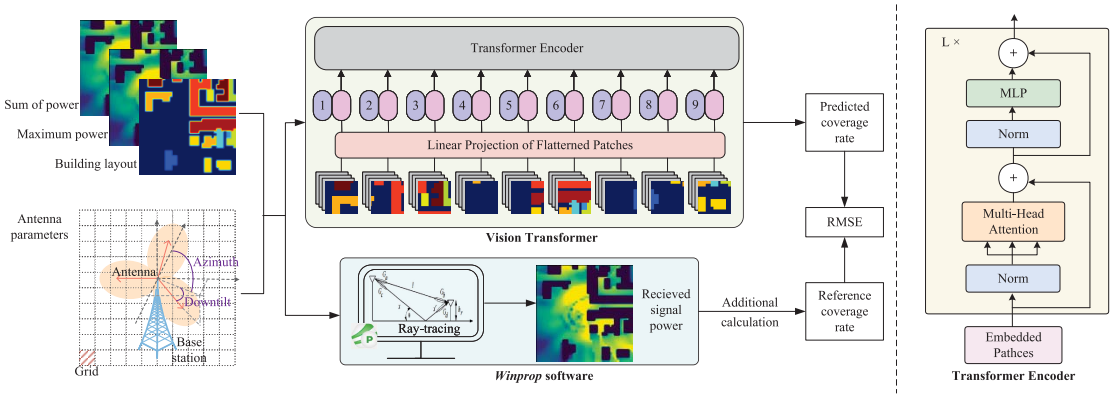

Fig. 1. Proposed coverage prediction workflow.

The output $\boldsymbol{y} = f_{\mathrm{LN}}(\boldsymbol{z}_{L,1})$ is the predicted coverage rate of ViT, where $\boldsymbol{z}_{L,1}$ is the first component of $\boldsymbol{z}_L$.

The standard self-attention mechanism, formalized as $f_{\mathrm{SA}}(\cdot)$, involves the following steps. First, the input $\boldsymbol{z}$ is projected into the query, key, and value representations via a shared linear transformation:

$$[\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] = \boldsymbol{z}\,\boldsymbol{U}, \tag{7}$$

where $\boldsymbol{U}$ is the concatenated weight matrix for all three projections. Next, the attention weights are computed using the scaled dot-product of the query and key vectors:

$$\boldsymbol{A} = softmax\left(\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}/\sqrt{d_k}\right), \tag{8}$$

where d_k is the dimensionality of the key vectors. Finally, the output of the self-attention layer is obtained by applying the attention weights to the value vectors:

$$f_{\mathrm{SA}}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{v}. \tag{9}$$

Multi-head self-attention is an extension of self-attention, in which m self-attention operations are performed in parallel, called multi-head, and different heads focus on different features:

$$f_{\mathrm{MSA}}(\boldsymbol{z}) = \left[f_{\mathrm{SA}_1}(\boldsymbol{z}); f_{\mathrm{SA}_2}(\boldsymbol{z}); \cdots; f_{\mathrm{SA}_m}(\boldsymbol{z})\right] \boldsymbol{U}_m. \tag{10}$$

A complete overview of the coverage prediction process is presented in Fig. 1.

IV. Numerical Results

We evaluate coverage prediction error and computational efficiency on 200m × 200m urban terrain maps, discretized into 5m × 5m grids for a balance between resolution and computational cost. Each map contains a 30m-high mmWave base station placed at the center, transmitting at 15.4dBm and 30GHz. Antenna configurations are systematically evaluated across 36 azimuth angles from 0° to 350° in 10° increments and 7 elevation downtilt angles from 0° to 30° in 5° steps, ensuring comprehensive spatial sampling of the radiation profile.

To assess model generalizability, the dataset is split into training and test sets with ratios from 10%:90% to 90%:10% in 10% increments. Models are trained in a supervised manner to minimize the root mean squared error (RMSE) between predicted and reference coverage rates. We adopt the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}, and a batch size of 256, while applying a dropout rate of 0.05. The model is considered to have converged and training is stopped when the test set RMSE exhibits oscillations within ± 0.001 over 50 consecutive epochs. Neural network training are implemented in PyTorch and executed on a server equipped with eight NVIDIA RTX A6000 GPUs. Furthermore, to investigate the impact of model capacity on prediction performance, we adopt three ViT architectures of increasing scale: ViT-Base, ViT-Large, and ViT-Huge. These models vary in the number of transformer layers, hidden dimensions, feed-forward (MLP) size, and attention heads, as detailed in Table I.

TABLE I
Parameters of ViT Models

Model	Layers	Hidden size	MLP size	Heads
ViT-Base	12	768	3072	12
ViT-Large	24	1024	4096	16
ViT-Huge	32	1280	5120	20

We compare the coverage prediction error of the traditional COST 231 propagation model [14] and two classical CNN architectures, AlexNet [15] and ResNet [16], as well as a recent CNN-based model [17], which we refer to as PL-CNN for convenience, that uses map-based data for propagation estimation, all against our ViT models. The COST 231 model exhibits the highest RMSE at approximately 10.8%, which is significantly larger than all learning-based methods. Fig. 2 presents the RMSE results for the neural network models. All models show a consistent downward trend in prediction error as the training data increases. Among CNN baselines, the RMSEs are 7.5% and 4.5% for AlexNet at 10% and 90% training data, 5.4% and 2.2% for ResNet, and 6.7% and 3.2% for PL-CNN, respectively. In contrast, all ViT models achieve lower prediction error across all data regimes. Specifically, ViT-Huge achieves the best performance, with RMSE decreasing from about 3% using 10% of training data to below 1.2% when trained on 90% of the dataset. ViT-Large follows closely with slightly higher error rates, and ViT-Base

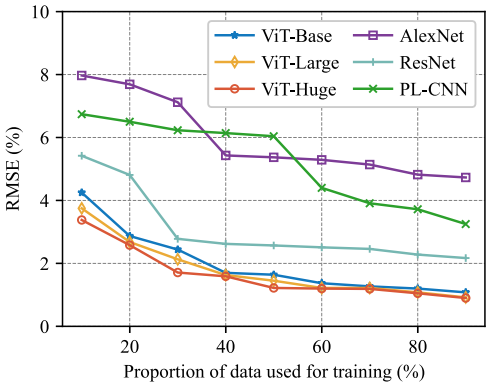

Fig. 2. Evaluation of ViT models and CNNs in mmWave coverage prediction with varying training splits.

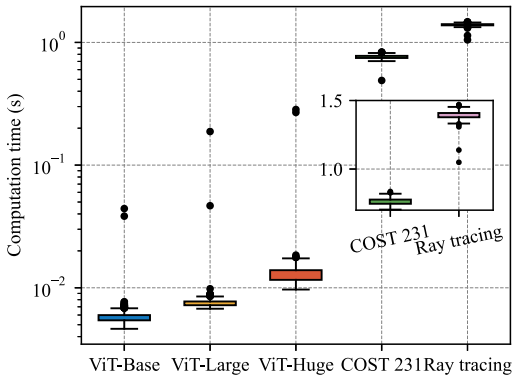

Fig. 3. Computation time comparison across coverage prediction methods.

maintains an RMSE under 2% once the training set exceeds 40%. These results demonstrate the superior capability of the transformer-based architectures to capture spatial correlations and deliver highly accurate coverage predictions.

To further evaluate the computation efficiency of the proposed method, we compare the average runtime of three ViT models with two traditional methods: the COST 231 empirical model and the ray-tracing model. Each test sample is executed 10 times, and the results are averaged. The overall average runtime and the standard deviation of per-sample averages are then computed across all samples. All experiments are conducted on the same hardware platform. As shown in Fig. 3, the ViT models achieve much faster computation. In particular, ViT-Base achieves an average runtime of only 5.8ms, which is substantially lower than that of the COST 231 model (0.76s) and the ray-tracing method (1.39s). These results demonstrate that the proposed approach offers not only high predictive performance but also notable computation efficiency, making it well-suited for latency-sensitive edge deployment scenarios.

V. Conclusion

In this letter, we developed an end-to-end ViT-based framework that encodes environmental features-including building distribution, interference maps, and antenna deployment-into image-like representations to directly predict coverage rate, without relying on traditional propagation models with explicit path-loss computation, thereby avoiding intermediate modeling errors. Experimental results demonstrate that our framework achieves lower prediction error, with ViT-Huge reducing it to sub-1% RMSE compared with ray tracing and consistently outperforming both empirical propagation models and CNN baselines. At the same time, it provides much faster inference, running up to two orders of magnitude quicker than propagation-based approaches, which is crucial for latency-sensitive edge scenarios. This balance of accuracy and efficiency makes the approach highly effective for edge network optimization. Future work entails hyperparameter tuning, memory and energy efficiency, scalability to larger and diverse environments, and practical update mechanisms for integration into network management systems.

References

[1] X. Yu, J. Zhang, M. Haenggi, and K. B. Letaief, "Coverage analysis for millimeter wave networks: The impact of directional antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1498–1512, Jul. 2017.

[2] Y. Yang and S. Wang, "EdgeOPT: A competitive algorithm for online parallel task scheduling with latency guarantee in mobile edge computing," *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 7077–7092, Nov. 2024.

[3] L. Shen, Y. Zhang, and S. Wang, "Codebook based antenna configuration: A new network planning paradigm for mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10368–10379, Aug. 2023.

[4] Y. Wu, X. Zhang, J. Ren, H. Xing, Y. Shen, and S. Cui, "Latency-aware resource allocation for mobile edge generation and computing via deep reinforcement learning," *IEEE Netw. Lett.*, vol. 6, no. 4, pp. 237–241, Dec. 2024.

[5] M. Iskander and Z. Yun, "Propagation prediction models for wireless communication systems," *IEEE Trans. Microw. Theory Techn.*, vol. 50, no. 3, pp. 662–673, Mar. 2002.

[6] Y. Okumura et al., "Field strength and its variability in VHF and UHF land-mobile radio service," *Rev. Electr. Comm. Lab.*, vol. 16, no. 9, pp. 825–873, Sept. 1968.

[7] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Technol.*, vol. 29, no. 3, pp. 317–325, Aug. 1980.

[8] H. Ling, R. Chou, and S. Lee, "Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity," *IEEE Trans. Antennas Propag.*, vol. 37, no. 2, pp. 194–205, Feb. 1989.

[9] G. Feng, J. Huang, and H. Su, "A new ray tracing method based on piecewise conformal transformations," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 4, pp. 2040–2052, Apr. 2022.

[10] S. Mohammadjafari, S. Roginsky, E. Kavurmacioglu, M. Cevik, J. Ethier, and A. B. Bener, "Machine learning-based radio coverage prediction in urban environments," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 4, pp. 2117–2130, Dec. 2020.

[11] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. ICLR*, May 2021, pp. 1–22.

[12] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[13] R. Hoppe, G. Wölfle, and U. Jakobus, "Wave propagation and radio network planning software WinProp added to the electromagnetic solver package FEKO," in *Proc. ACES*, Florence, Italy, Mar. 2017, pp. 1–2.

[14] E. Damosso and L. Correia, *Cost Action 231: Digital Mobile Radio Towards Future Generation Systems: Final Report*, Eur. Comm., Brussels, Belgium, 1999,

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1–9.

[16] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 1–12.

[17] R. Dempsey, J. Ethier, and H. Yanikomeroglu, "Map-based path loss prediction in multiple cities using convolutional neural networks," *IEEE Antennas Wireless Propag. Lett.*, vol. 24, pp. 1989–1993, 2025.