Liquan CHEN, Zixuan YANG, Peng ZHANG, Yang MA, 2025. Efficient privacy-preserving scheme for secure neural network inference. *Frontiers of Information Technology & Electronic Engineering*, 26(9):1609-1623.
https://doi.org/10.1631/FITEE.2400371

Efficient privacy-preserving scheme for secure neural network inference

Key words: Secure neural network inference; Convolutional neural network; Privacy-preserving; Homomorphic encryption; Secret sharing

Corresponding author: Liquan CHEN
E-mail: Lqchen@seu.edu.cn
ORCID: https://orcid.org/0000-0002-7202-4939

Motivation

Targeting the inadequacy of inefficiency and high communication overhead in existing encrypted image inference schemes, an efficient privacy-preserving scheme for secure neural network inference is proposed which ensures the privacy of both users and cloud servers while achieving fast and accurate ciphertext inference.

Main idea

- A network parameter merging approach is proposed, targeting consecutive multi-layer linear operations. This approach merges the parameters of multiple layers from the preprocessing stage and the online stage into a single layer for computation, effectively reducing the communication overhead and inference time.
- A fast convolution algorithm is proposed to address the heavy convolutional computations in the preprocessing stage. This method transforms homomorphic convolutional calculations into homomorphic matrix–vector multiplication calculations, adopting diagonal schemes for convolutional calculations, significantly reducing the computational time in the preprocessing stage.
- The efficiency of the proposed scheme is validated by the experimental analysis. Compared to the prior works, our scheme reduces both the computational time and communication overhead.

Method

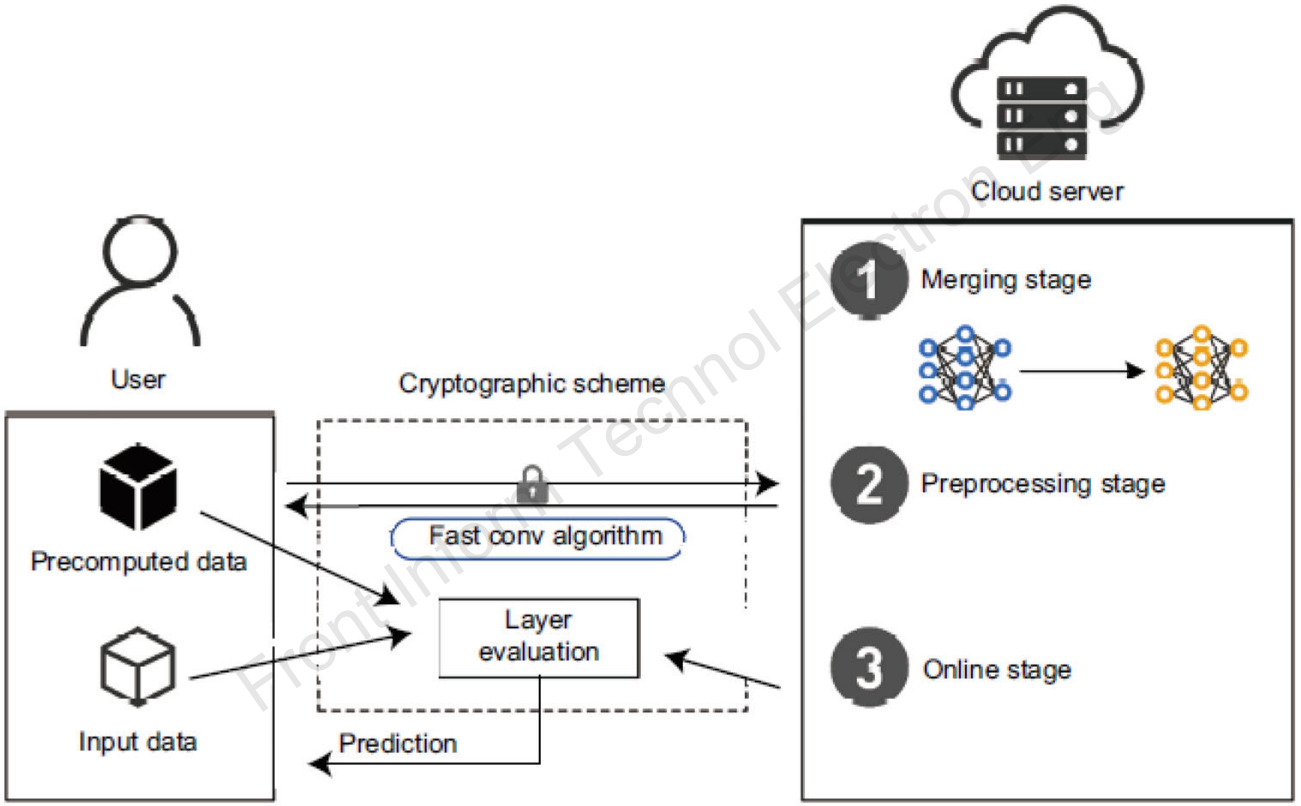

Fig. 4 System framework (conv: convolution)

Method

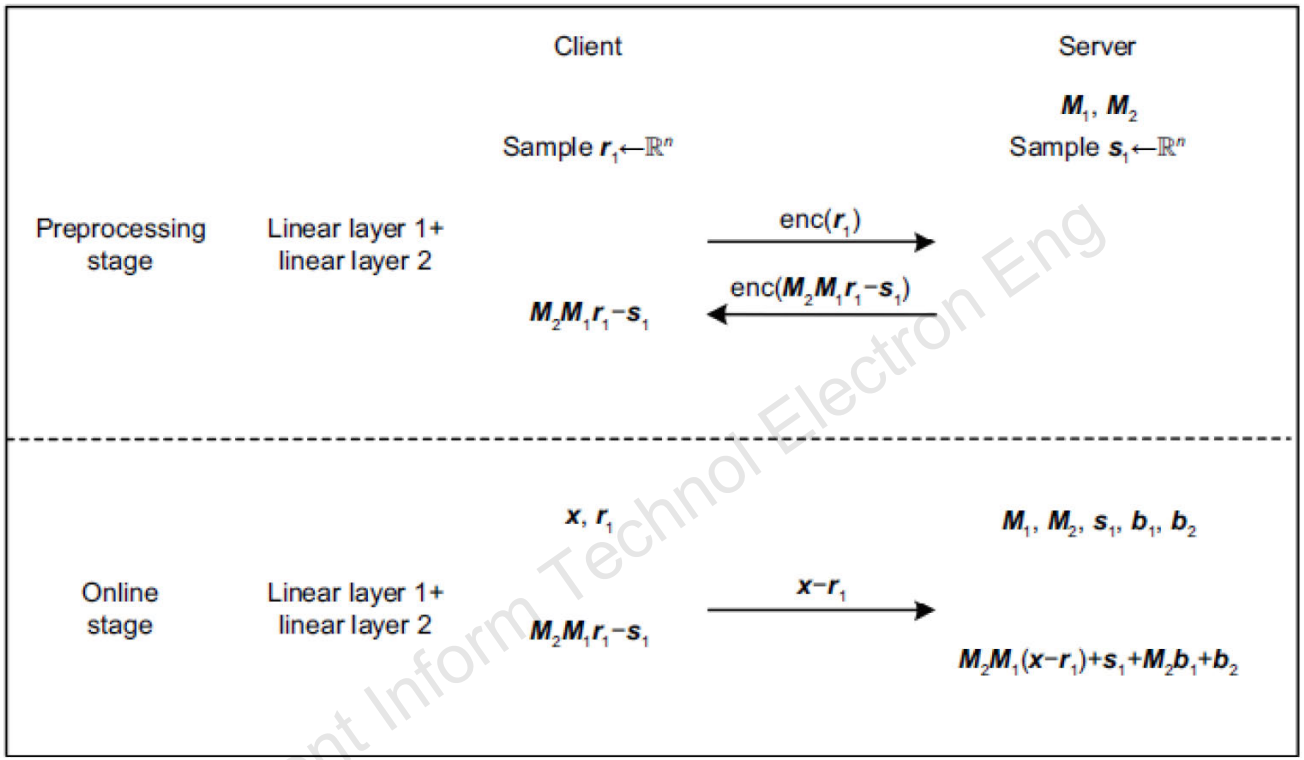

Fig. 5 Inference scheme of merging two consecutive linear layers

Table 2 Multiplication and addition operations in two consecutive linear layers

Approach	Inference result in the preprocessing stage	C–P operation	Inference result in the online stage	Floating-point operation
Nonmerging approach (DELPHI)	$M_1r_1 - s_1$; $M_2r_2 - s_2$	2 Multi Op; 2 Add Op	$M_1(x_1 - r_1) + (b_1 + s_1)$; $M_2(x_2 - r_2) + (b_2 + s_2)$	2 Multi Op; 2 Add Op
Merging approach	$M_2M_1r_1 - s_1$	1 Multi Op; 1 Add Op	$(M_2M_1)(x - r_1)$ $+(s_1 + M_2b_1 + b_2)$	1 Multi Op; 1 Add Op

Method

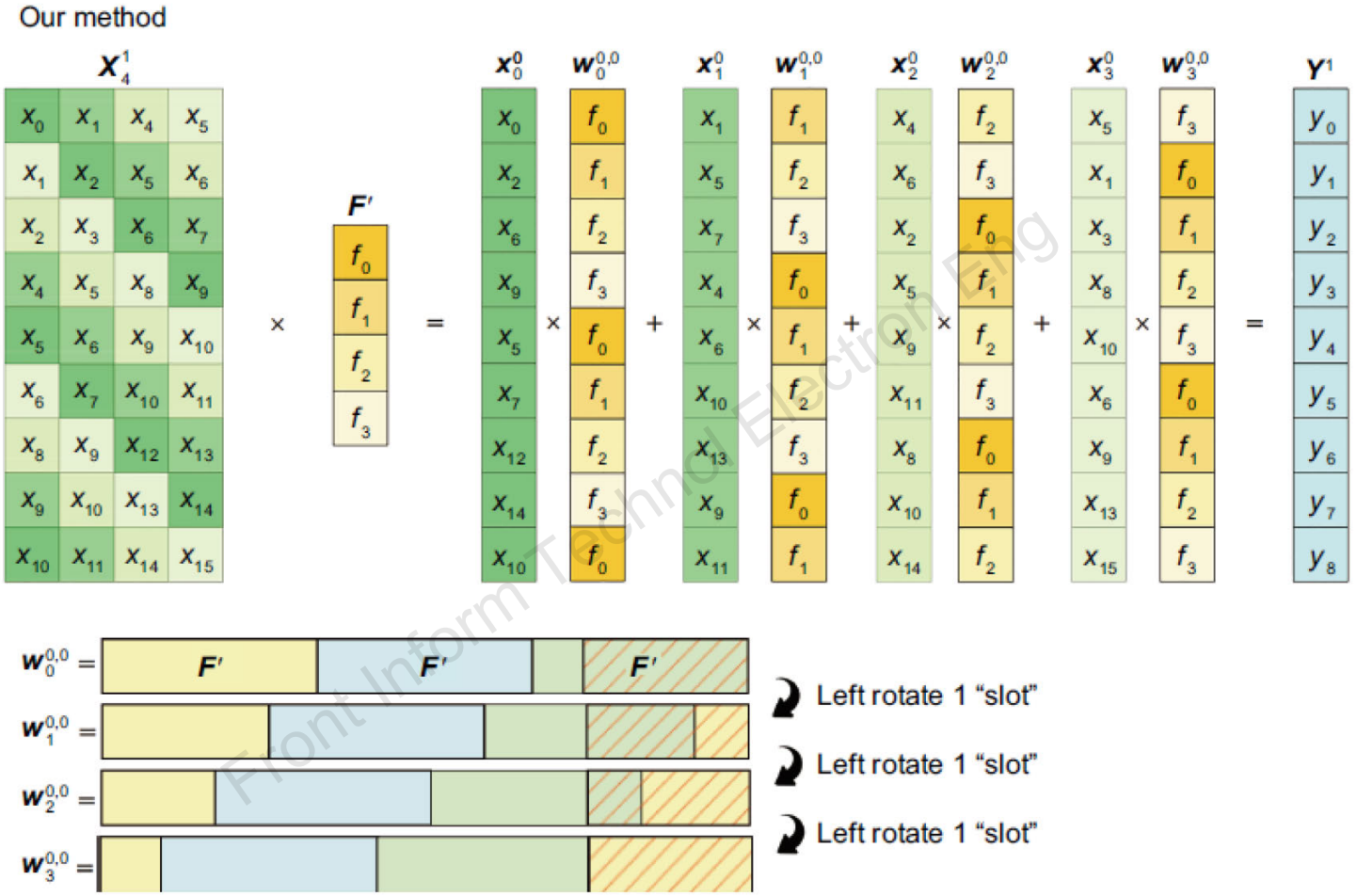

Fig. 8 Kernel processing and computational steps in the fast convolution algorithm

The i^{th} element of $\boldsymbol{x}_t^0$ corresponds to the i^{th} row and j^{th} column element of $\boldsymbol{X}_{f_{\text{Size}}}^1$, where $t = (j - (i \bmod f_{\text{Size}})) \bmod f_{\text{Size}}$.

Method

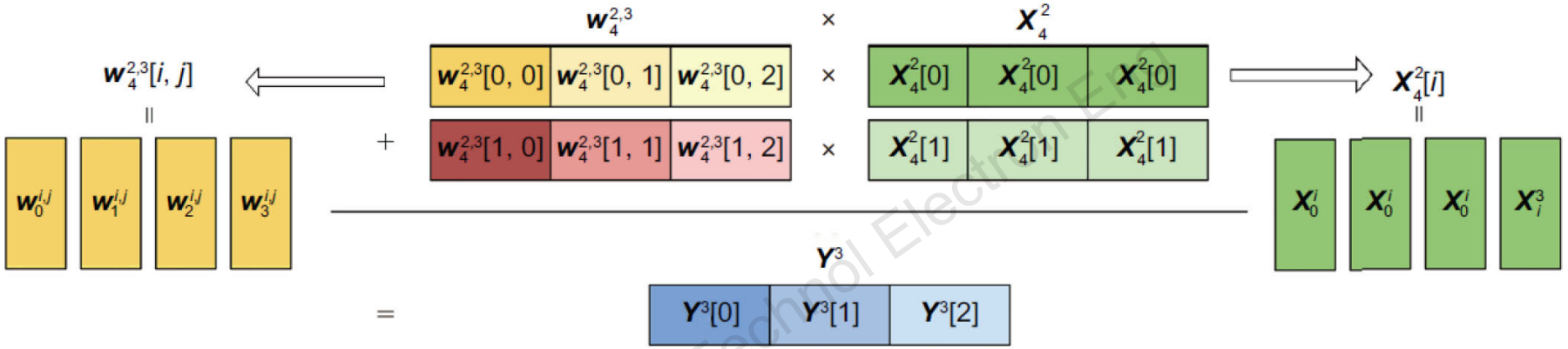

Fig. 9 Multidimensional fast convolution

We placed multiple $W_{f_{\text{Size}}}^{\text{IC,OC}}[i,j]$ corresponding to the same input dimension *i* but different output dimensions *j* into the same plaintext. Moreover, we replicated $X_{f_{\text{Size}}}^{\text{IC}}[i]$ with the same input dimension *i* multiple times and placed them into the same ciphertext for computation, aiming to reduce the number of computations for multi-dimensional convolutions and consequently decrease the computational time.

Results

Table 3 Network description

Layer	Description
Conv I	Input size: $28 \times 28 \times 1$; kernel size: $5 \times 5 \times 1$; stride: (1, 1); number of filters: 16; output size: $24 \times 24 \times 16$
Activation I	Calculates ReLU for each input
Maxpool I	Pool size: (2, 2); stride: (2, 2)
Conv II	Input size: $12 \times 12 \times 16$; kernel size: $5 \times 5 \times 16$; stride: (1, 1); number of filters: 16; output size: $8 \times 8 \times 16$
BN	Input size: $8 \times 8 \times 16$; output size: $8 \times 8 \times 16$
Activation II	Calculates ReLU for each input
Maxpool II	Pool size: (2, 2); stride: (2, 2)
Flatten	Multidimensional arrays are expanded to one-dimensional arrays
FC I	Fully connects the incoming 256 nodes to the outgoing 100 nodes
Activation III	Calculates ReLU for each input
FC II	Fully connects the incoming 100 nodes to the outgoing 10 nodes
Softmax	

Table 4 Inference accuracy of different models

Model	Accuracy (%)	
	MINST	Fashion-MNIST
Ours	99.24	90.26
CryptoNets (Dowlin et al., 2016)	98.95	87.27
GAZELLE (Juvekar et al., 2018)	99.08	90.26
DELPHI (Mishra et al., 2020)	99.08	90.26
Approximated Mish (Yagyu et al., 2023)	99.30	84.68
Approximated SiLU (Lai et al., 2024)	97.17	89.35

Results

Table 5 Comparison of the inference time based on the MNIST dataset

Scheme	Time in the preprocessing stage (ms)		Time in the online stage (ms)	
	Linear	Nonlinear	Linear	Nonlinear
Ours	2059.10	1541.34	1.39	894.52
DELPHI (Mishra et al., 2020)	2089.51	1541.34	1.56	894.52
GAZELLE (Juvekar et al., 2018)			2089.51	2657.64
CryptoNets (Dowlin et al., 2016)			264 970.28	

Table 6 Comparison of the inference time based on the Fashion-MNIST dataset

Scheme	Time in the preprocessing stage (ms)		Time in the online stage (ms)	
	Linear	Nonlinear	Linear	Nonlinear
Ours	2406.4	1789.82	1.47	1038.73
DELPHI (Mishra et al., 2020)	2411	1789.82	1.67	1038.73
GAZELLE (Juvekar et al., 2018)			2411	3066.56
CryptoNets (Dowlin et al., 2016)			315 674.1	

Results

Table 7 Description of convolutional layers in the experiment

Group	Input size	Kernel size	Stride	Number of filters	Padding
1	$32 \times 32 \times 8$	5×5	(1, 1)	64	Valid
2	$78 \times 78 \times 16$	3×3	(2, 2)	48	Same
3	$128 \times 128 \times 8$	5×5	(2, 2)	16	Valid
4	$64 \times 64 \times 16$	3×3	(1, 1)	24	Same
5	$8 \times 8 \times 6$	3×3	(1, 1)	4	Same
6	$16 \times 16 \times 4$	5×5	(1, 1)	2	Valid
7	$24 \times 24 \times 3$	9×9	(2, 2)	4	Valid
8	$32 \times 32 \times 3$	7×7	(2, 2)	2	Same

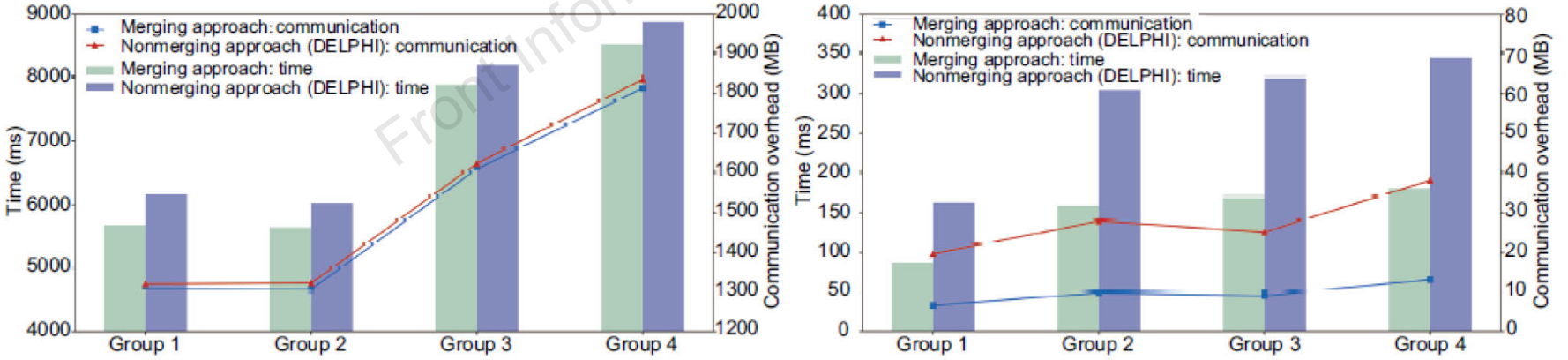

Fig. 10 Comparison of efficiency between merging and nonmerging approaches in the preprocessing stage (a) and online stage (b)

Results

Table 8 Computing time comparison of merging approaches

Method	Computing time (ms)			
	Group 5	Group 6	Group 7	Group 8
Ours	651.2	1184.5	2945.5	1762.6
Matrix method	3075.6	3490.2	3118.7	6117.8

Table 9 Computing time comparison of convolution algorithms

Method	Computing time (s)			
	Convolution I	Convolution II	Convolution III	Convolution IV
Ours	13.58	23.57	4.13	1.93
Improved naive	129.55	615.61	42.95	21.66
Packed SC (Juvekar et al., 2018)	13.93	24.84	4.14	1.94

Results

Four convolutional layer configurations for testing:

Convolution I: padding: same; stride: (2, 2); input size: $228 \times 228 \times 3$; kernel size: $3 \times 3 \times 3$; number of filters: 64.

Convolution II: padding: same; stride: (2, 2); input size: $64 \times 64 \times 32$; kernel size: $3 \times 3 \times 32$; number of filters: 128.

Convolution III: padding: valid; stride: (2, 2); input size: $128 \times 128 \times 3$; kernel size: $3 \times 3 \times 3$; number of filters: 64.

Convolution IV: padding: valid; stride: (2, 2); input size: $32 \times 32 \times 16$; kernel size: $5 \times 5 \times 16$; number of filters: 24.

Conclusions

This study presents an efficient privacy-preserving secure neural network inference scheme, in which we propose an approach of merging network parameters for consecutive linear layers and a fast convolution algorithm to reduce the computational time of linear layers in the preprocessing stage. Our experiments demonstrate that our scheme provides superior performance in secure neural network inference under the premise of protecting privacy for both parties.

Liquan CHEN received the Ph.D. degree from Southeast University, China, in 2005. He worked as a postdoc in Southeast University from 2005 to 2007, and an associate professor at Southeast University from 2008 to 2018. He worked as a visiting scholar in the National University of Singapore, Singapore, from 2011 to 2012. He is now a professor at the School of Cyber Science and Engineering, Southeast University, Nanjing, China. His research interests include information security, cryptography, and network security protocol.

Zixuan YANG received the B.S. degree in cyber science and engineering and the M.S. degree in network and information security from Southeast University, Nanjing, China, in 2022 and 2025, respectively. Her research interests include security image retrieval.

Peng ZHANG received the B.S. and M.S. degrees in cyber science and engineering from Southeast University, Nanjing, China, in 2022 and 2025, respectively. His research interests include security image retrieval.

Yang MA is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Southeast University, Nanjing, China. He is serving as the Director of the Information and Communication Management Division, Jiangsu Provincial Communication Administration. Holding the title of senior engineer, he is recognized as a High-Level Talent under the Jiangsu Provincial 333 Program. He has participated in and led the completion of five projects under the National Information Security Project Plan. He has published over 10 papers in core domestic and international journals and has obtained authorization for seven national patents. Additionally, he has twice been awarded the Science and Technology Progress Award by the China Institute of Communications for projects he has led or participated in.