T-TransNet: Ternary Attention Network for CSI Feedback in FDD Massive MIMO System

Seongjin Hwang[1,*], Seungmin Choi[1,*], and Hyun Jong Yang[1,2]
[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2]Institute of New Media and Communications, Seoul National University, Seoul, South Korea
{sjh1753, seungminchoi22, hjyang}@snu.ac.kr

***Abstract*—Channel state information (CSI) feedback is critical for frequency division duplexing (FDD) massive multiple-input multiple-output (MIMO) systems, yet the dimensionality of modern antenna–subcarrier configurations makes low-latency uplink reporting challenging. Deep learning (DL) based autoencoder schemes have proven effective for compressing and reconstructing CSI, and attention-based variants such as TransNet report notable reconstruction quality across a range of compression ratios. However, these gains come at the expense of substantial computational and memory footprints, which limit real-time deployment at user equipment (UE). We address this gap with *T-TransNet*, a ternary attention network that applies trainable ternary quantization to the fully connected (FC) projections inside multihead attention blocks, thereby reducing UE-side arithmetic. Building on insights from binary designs for CSI feedback and from trained ternary quantization methods in computer vision, we introduce two complementary techniques that preserve accuracy under aggressive quantization: (i) a *Learnable GLAQ* activation that nonlinearly compresses outlier magnitudes with a learnable scale; and (ii) *column-wise ternary quantization* that allocates independent scaling to each output channel of FC weights.**

***Index Terms*—CSI feedback, massive MIMO, FDD, attention network, ternary quantization, lightweight encoder.**

I. Introduction

Massive MIMO is a cornerstone physical-layer technology for 5G and beyond because large antenna arrays enable pronounced spectral and energy-efficiency gains when accurate downlink CSI is available at the base station (BS) for precoding and scheduling. In FDD operation, however, the downlink channel must be estimated at the user equipment (UE) and fed back to the BS, and the resulting overhead scales with the product of antenna count and subcarriers, quickly becoming prohibitive in large-array regimes. Early work demonstrated that learned autoencoder architectures (CsiNet) can exploit channel structure more effectively than compressed-sensing approaches at practical compression ratios, enabling high-quality CSI reconstruction even when classical sparsity assumptions are imperfect [1].

More recently, attention-based architectures have been explored for CSI feedback. Drawing inspiration from the Transformer, TransNet applies full self-attention within an encoder–decoder pipeline and reports consistent reconstruction gains over convolutional baselines across several compression scales [2]. Attention excels at capturing long-range dependencies across the angular–delay structured CSI matrix, but introduces sizable FC projections per attention head, increasing UE-side memory and computation.

In practice, UE hardware and uplink latency constraints motivate lightweight encoders whose inference pathways exhibit low arithmetic intensity, modest parameter storage, and reduced memory bandwidth. In [3], [4], binary quantized CSI encoders show that aggressive quantization, particularly within large FC bottlenecks, can reduce UE complexity while preserving competitive reconstruction.

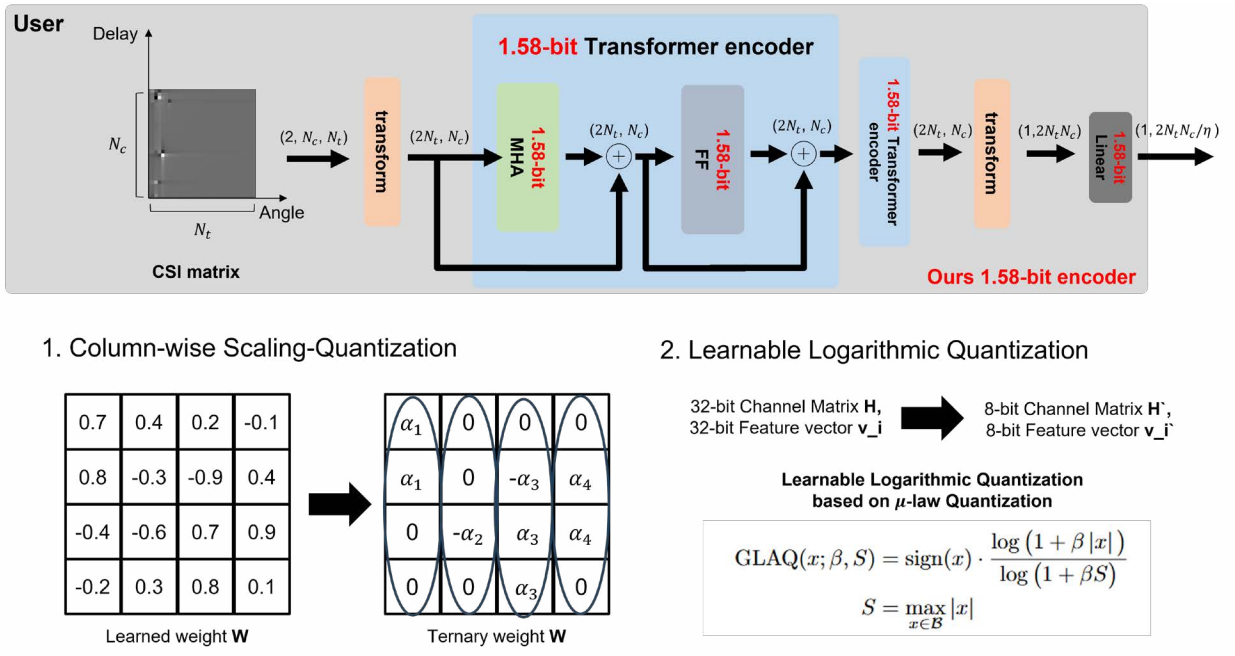
Moving beyond 1-bit representations, ternary weight networks (TWN) and trained ternary quantization (TTQ) methods in the broader DL literature strike a more favorable accuracy, efficiency balance by introducing a zero state and learnable asymmetric scaling for positive and negative weights [5], [6]. These insights suggest that applying structured ternary quantization to the heavy FC components of attention blocks could yield a practical attention-based CSI encoder.

Our Contributions: Motivated by the above, we propose *T-TransNet*, a ternary-quantized attention architecture for CSI feedback in FDD massive MIMO systems. Our main contributions are as follows.

- We introduce a variant oriented to UE of TransNet in which the dense projections of multi-head attention (MHA) and output mixing layers are *ternarized* to $\{-\alpha, 0, \alpha\}$ with a learnable scaling.
- We develop a *Learnable Generalized Logarithmic Activation Quantizer (GLAQ)* that adaptively compresses activation dynamic range prior to quantization, mitigating information loss from low-bit weights.
- We propose *column-wise ternary quantization* that assigns independent scaling factors to each output column of an FC weight matrix, improving the representational capacity under 2-bit storage.

The remainder of this paper is organized as follows. Section II reviews the FDD CSI feedback model and formalizes the learning objective. Section III describes the T-TransNet architecture, the ternary quantization procedure, the activation of Learnable GLAQ and the training algorithm. Section IV (to be added) reports numerical results. Section V concludes.

*: These first authors have equally contributed.
Corresponding Author: Hyun Jong Yang



$$\mathrm{GLAQ}(x;\beta,S) = \mathrm{sign}(x)\cdot\frac{\log\left(1+\beta\,|x|\right)}{\log\left(1+\beta S\right)}$$
$$S = \max_{x\in\mathcal{B}}|x|$$

Fig. 1. Proposed T-TransNet architecture. UE-side encoder applies (1) Learnable GLAQ activation; (2) column-wise trained ternary quantization within multihead attention projections; (3) optional feature quantization for uplink bits. BS-side decoder mirrors TransNet with higher-capacity refinement. 1.58-bit means ternary quantized value.

II. System Model and Problem Formulation

We consider a single-cell downlink FDD massive MIMO orthogonal frequency-division multiplexing (OFDM) system with N_t BS transmit antennas and a single UE receive antenna ($N_r = 1$). Let $\tilde{N}_c$ denote the total number of OFDM subcarriers. The received symbol on subcarrier n is

$$y_n = \tilde{\mathbf{h}}_n^{\mathrm{H}}\mathbf{v}_n x_n + z_n, \quad n = 1,\ldots,\tilde{N}_c, \tag{1}$$

where $\tilde{\mathbf{h}}_n \in \mathbb{C}^{N_t\times 1}$ is the downlink channel vector, $\mathbf{v}_n$ the precoder, x_n the transmitted data symbol, and z_n additive noise. Stack the spatial-frequency domain channel vectors to form $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1,\ldots,\tilde{\mathbf{h}}_{\tilde{N}_c}]^{\mathrm{H}} \in \mathbb{C}^{\tilde{N}_c\times N_t}$.

A. Angular–Delay Domain Sparsification

Following standard practice, we apply a 2D discrete Fourier transform (DFT) across subcarriers and antenna indices to expose angular–delay structure:

$$\mathbf{H} = \mathbf{F}_c\tilde{\mathbf{H}}\mathbf{F}_t^{\mathrm{H}}, \tag{2}$$

where $\mathbf{F}_c \in \mathbb{C}^{\tilde{N}_c\times\tilde{N}_c}$ and $\mathbf{F}_t \in \mathbb{C}^{N_t\times N_t}$ are unitary DFT matrices over frequency and antenna dimensions, respectively. Because multipath delay spread is limited, significant energy concentrates in the first N_a delay taps of $\mathbf{H}$. We therefore truncate $\mathbf{H}$ to its leading rows N_a to obtain $\mathbf{H}_a \in \mathbb{C}^{N_a\times N_t}$, reducing the dimensionality with the channel information.

B. Learning-Based Compression and Recovery

Let $f_C(\cdot;\Theta_C)$ and $f_R(\cdot;\Theta_R)$ denote the UE-side encoder and BS-side decoder neural networks, respectively. The encoder maps the (vectorized) truncated channel to a length-M latent vector $\mathbf{v} \in \mathbb{R}^M$ that is fed back over the uplink:

$$\mathbf{v} = f_C(\mathbf{H}_a;\Theta_C), \qquad M = \eta N_a N_t \times 2, \tag{3}$$

where $\eta \in (0,1]$ is the compression ratio and the factor of two accounts for separate real and imaginary parts. After ideal uplink transfer, the BS reconstructs

$$\widehat{\mathbf{H}}_a = f_R(\mathbf{v};\Theta_R), \tag{4}$$

from which the full spatial-frequency domain estimate $\widehat{\tilde{\mathbf{H}}}$ is recovered by zero-padding and inverse transforms:

$$\widehat{\tilde{\mathbf{H}}} = \mathbf{F}_c^{\mathrm{H}}\begin{bmatrix}\widehat{\mathbf{H}}_a & \mathbf{0}\end{bmatrix}\mathbf{F}_t. \tag{5}$$

C. Training Objective

The parameters (Θ_C,Θ_R) are learned by minimizing a distortion metric between $\mathbf{H}_a$ and its reconstruction. We adopt mean-squared error (MSE) or normalized MSE (NMSE) in a training set $\mathcal{D}$:

$$\min_{\Theta_C,\Theta_R}\ \mathbb{E}_{\mathbf{H}_a\sim\mathcal{D}}\left[\,\|\mathbf{H}_a - f_R(f_C(\mathbf{H}_a;\Theta_C);\Theta_R)\|_2^2\,\right] \tag{6}$$

III. Proposed T-TransNet Architecture

In this paper, we propose T-TransNet which described in 1.

A. Design Overview With Transformer Encoder

T-TransNet inherits the encoder, decoder attention backbone of TransNet [2] but introduces UE-side quantization and activation compression motivated by the structure of attention network. For completeness, consider the sequence of input features in an MHA block as $\mathbf{X} \in \mathbb{R}^{L\times D}$. With N_h heads

and $d_h = D/N_h$, the k-th head forms query, key, and value projections

$$\mathbf{Q}_k = \mathbf{X}\mathbf{W}_k^Q \in \mathbb{R}^{L\times d_h}, \tag{7a}$$

$$\mathbf{K}_k = \mathbf{X}\mathbf{W}_k^K \in \mathbb{R}^{L\times d_h}, \tag{7b}$$

$$\mathbf{V}_k = \mathbf{X}\mathbf{W}_k^V \in \mathbb{R}^{L\times d_h}, \tag{7c}$$

where $\mathbf{W}_k^Q, \mathbf{W}_k^K, \mathbf{W}_k^V$ are dense projection matrices. Scaled dot-product attention computes similarity scores and attention weights as

$$\mathbf{M}_k = \mathrm{softmax}\left(\frac{\mathbf{Q}_k\mathbf{K}_k^\mathsf{T}}{\sqrt{d_h}}\right) \in \mathbb{R}^{L\times L}, \tag{8}$$

leading to the per-head output

$$\mathbf{A}_k = \mathbf{M}_k\mathbf{V}_k \in \mathbb{R}^{L\times d_h}. \tag{9}$$

The head outputs are concatenated and mixed linearly by an output projection $\mathbf{W}^O$:

$$\mathbf{A} = \mathrm{Concat}(\mathbf{A}_1, \ldots, \mathbf{A}_{N_h})\mathbf{W}^O. \tag{10}$$

B. Ternary Quantization of Attention Projections

We convert linear project in III-A to ternary quantization weight. Consider an linear matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ in an MHA projection and an output projection. Our goal is to approximate $\tilde{\mathbf{W}}_{\text{ternary}} = \alpha\mathbf{W}_{\text{signed}}$, where $\mathbf{W}_{\text{signed}} \in \{+1, 0, -1\}^{m\times n}$ are binary masks denoting positive and negative assignments and scale factor $\alpha \in \mathbb{R}$. We begin with symmetric scaling $\alpha > 0$ and define

$$[\tilde{\mathbf{W}}_{\text{ternary}}]_{ij} = \begin{cases} +\alpha, & w_{ij} > \Delta, \\ 0, & |w_{ij}| \le \Delta, \\ -\alpha, & w_{ij} < -\Delta, \end{cases} \tag{11}$$

where threshold Δ and scale α are chosen to minimize $\|\mathbf{W} - \alpha\mathbf{W}_{\text{signed}}\|$, as in ternary weight networks (TWN) [5], [6]. A common closed-form heuristic is $\Delta = 0.7 \cdot \mathbb{E}[|w_{ij}|]$.

C. Column-Wise Ternary Quantization

CSI matrices exhibit direction-dependent energy across antenna dimensions; likewise, attention projections feeding different heads or feature channels can have markedly different dynamic ranges. We therefore generalize TTQ to *column-wise* scaling: for output column j we learn $(\alpha_j^+, \alpha_j^-, \Delta_j)$. Quantization follows with column-specific thresholds and scales. At inference, multiplications reduce to additions/subtractions gated by ternary masks; column-specific scales can be folded into subsequent normalization layers, minimizing cost.

D. Gradient for Ternary Quantization.

Trained ternary quantization (TTQ) introduces α and ternary weight W α^- per layer and updates them via gradient descent [5]. Let $\mathcal{I}_j^+ = \{i : w_{ij} > \Delta_j\}$ and $\mathcal{I}_j^- = \{i : w_{ij} < -\Delta_j\}$. Given loss $\mathcal{L}$, gradients accumulate as

$$\frac{\partial\mathcal{L}}{\partial\alpha_j^+} = \sum_{i\in\mathcal{I}_j^+}\frac{\partial\mathcal{L}}{\partial\tilde{w}_{ij}}, \qquad \frac{\partial\mathcal{L}}{\partial\alpha_j^-} = \sum_{i\in\mathcal{I}_j^-}\frac{\partial\mathcal{L}}{\partial\tilde{w}_{ij}}, \tag{12}$$

Algorithm 1 Mini-batch Training Procedure for T-TransNet

Require: Batch of truncated channels $\{\mathbf{H}_a^{(b)}\}_{b=1}^B$; learning rates η_w, η_α

1: Stack real/imag parts and apply GLAQ activation (14)
2: **for** each attention projection matrix $\mathbf{W}$ **do**
3: Estimate per-column threshold $\Delta_j = t \cdot \mathbb{E}_i[|w_{ij}|]$ or use learnable Δ_j
4: Quantize: $\tilde{w}_{ij} \in \{-\alpha_j, 0, +\alpha_j\}$ via (11)
5: Forward propagate using ternary weights
6: **end for**
7: Compute loss $\mathcal{L}$ via NMSE (IV-A)
8: Backpropagate:
- Accumulate scale gradients $\partial\mathcal{L}/\partial\alpha_j^\pm$ via (12)
- Use STE gradient (13) for latent w_{ij}
- Use (16) for GLAQ α gradients

9: Update:
- $\alpha_j^\pm \leftarrow \alpha_j^\pm - \eta_\alpha \cdot \partial\mathcal{L}/\partial\alpha_j^\pm$
- $w_{ij} \leftarrow w_{ij} - \eta_w \cdot \partial\mathcal{L}/\partial w_{ij}$
- Update α in GLAQ with softplus parameterization

while gradients w.r.t. latent FP weights use scaled straight-through estimators (STEs):

$$\frac{\partial\mathcal{L}}{\partial w_{ij}} = \begin{cases} \alpha\frac{\partial\mathcal{L}}{\partial\tilde{w}_{ij}}, & w_{ij} > \Delta, \\ \frac{\partial\mathcal{L}}{\partial\tilde{w}_{ij}}, & |w_{ij}| \le \Delta, \\ -\alpha\frac{\partial\mathcal{L}}{\partial\tilde{w}_{ij}}, & w_{ij} < -\Delta. \end{cases} \tag{13}$$

This scheme learns both scaling and assignments during training and produces high-accuracy 2-bit models in large-scale vision tasks [6].

To counteract accuracy loss from low-bit weights, we introduce two mechanisms: (i) *Learnable GLAQ* activation, applied to the pre-projection activations to compress magnitude outliers; and (ii) *column-wise ternary quantization*, which extends layer-wise scaling to per-output-column scaling, capturing anisotropic channel statistics more faithfully.

A high-level block diagram is shown in Fig. 1.

E. Learnable GLAQ Activation

Let x denote an activation element (real-valued). We define the generalized logarithmic activation quantizer (GLAQ) as

$$\mathrm{GLAQ}(x;\beta,S) = \mathrm{sign}(x)\cdot\frac{\log\left(1+\beta\,|x|\right)}{\log\left(1+\beta S\right)}, \qquad S = \max_{x\in\mathcal{B}}|x| \tag{14}$$

with learnable scale parameter $\alpha > 0$ and batch (or running) normalization scale S. The mapping is odd, monotone in $|x|$, and compresses large magnitudes (logarithmic growth) into $[-1, 1]$. Small magnitudes remain approximately linear (for $\alpha\,|x| \ll 1$), while outliers are squashed, improving robustness to subsequent low-bit weight multiplication.

Gradients for GLAQ Activation: Ignoring the measure-zero non-differentiability at $x = 0$, the partial derivatives used in backpropagation are

$$\frac{\partial\,\mathrm{GLAQ}(x)}{\partial x} = \frac{\beta}{(1+\alpha|x|)\,\log(1+\beta S)} \tag{15}$$

$$\frac{\partial\,\mathrm{GLAQ}(x)}{\partial \beta} = \mathrm{sign}(x)\cdot\frac{f(x;\beta)}{[\log(1+\beta S)]^2}. \tag{16}$$

$$f(x;\beta) = \frac{|x|}{1+\beta|x|}\log(1+\beta S) - \log(1+\beta|x|)\cdot\frac{S}{1+\beta S} \tag{17}$$

In practice we clip $\partial\mathrm{GLAQ}(x)/\partial x$ to a finite range and treat S either as a per-mini-batch constant or as an exponential moving average for stability. Parameter β is constrained via $\beta = \mathrm{softplus}(\tilde{\beta})$ during training.

F. Training Algorithm

Algorithm 1 summarizes mini-batch stochastic gradient descent (SGD) for T-TransNet.

IV. Experimental Results

This section will report reconstruction NMSE, parameter counts, and encoder FLOPs for T-TransNet versus Transnet across indoor and outdoor COST2100 scenarios and multiple compression ratios $\eta \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$. Moreover, our experimental setup is followed by Section IV-A.

We deploy two variants of T-TransNet, namely T-TransNet-A and T-TransNet-B, which differ in the scope of quantization. T-TransNet-A applies quantization to all components, including the multi-head attention (MHA) layers and the final linear block. In contrast, T-TransNet-B restricts quantization to the final linear block, which carries the main computational burden due to the flattened vector representation.

Table I highlights the performance degradation observed in the T-TransNet-A model. However, Table II shows that T-TransNet-B achieves competitive results compared to the original TransNet.

A. Simulation Setup

- Dataset: COST2100 indoor and outdoor; split as in [1], [2].
- Channel dimensions: (N_t, N_a) = (32,32).
- Training: Adam, batch size sets 300, learning rates sets 0.3, and cosine based schedule.
- Metrics: NMSE (dB), encoder FLOPs.

$$\mathrm{NMSE} = \frac{\left\|\mathbf{H} - \widehat{\mathbf{H}}\right\|_2^2}{\|\mathbf{H}\|_2^2}.$$

TABLE I
NMSE Comparison of TransNet and T-TransNet-A

η	Method	NMSE indoor (dB)
1/4	TransNet	**-32.38**
	T-TransNet-A	-24.30

TABLE II
NMSE Comparison of TransNet and T-TransNet-B

η	Method	NMSE indoor (dB)	NMSE outdoor (dB)
1/4	TransNet	**-32.38**	-14.86
	T-TransNet-B	-29.40	**-15.03**
1/8	TransNet	-22.91	-9.99
	T-TransNet-B	**-24.02**	**-11.24**
1/16	TransNet	-15.00	**-7.82**
	T-TransNet-B	**-17.35**	-6.56
1/32	TransNet	-10.49	-4.13
	T-TransNet-B	**-11.83**	**-4.49**

*T-Transnet terminated at epoch 5000.

TABLE III
Comparision of Matrix Multiplication between 32-bit and column-wise ternary case.

Methods[a]	**# of 32-bit scalar mul.**	**# of bits**
32-bit	mnp	32mn
C-Ternary	mp	2mn + 32n
Ternary	0	2mn

B. Complexity Considerations

Assume an $m \times n$ FP matrix in attention projection. FP inference cost is mn MACs and mn parameters. A ternary matrix stores a 2-bit code per element plus scale(s). With layer-wise scaling, parameter storage reduces by $16\times$ relative to 32-bit float; with column-wise scaling, overhead adds $\mathcal{O}(n)$ scalars. Arithmetic reduces because multiplications by ± 1 become adds/subtracts and zeros skip; hardware implementations can further exploit sparsity when Δ_j induces many zeros. Section IV quantifies realized savings and UE latency.

In table II, we consider three matrix multiplication which AB, AC, and AD where a 32-bit matrix $A \in \mathbb{R}^{m\times n}$, a 32-bit matrix $B \in \mathbb{R}^{n\times p}$, a column-wise ternary matrix $C \in [\{-\alpha_j, 0, \alpha_j\}^n]_{j=1}^p$, and a ternary matrix $D \in \{-1, 0, 1\}^{n\times p}$

V. Conclusion

We presented T-TransNet, a ternary attention network that targets UE-efficient CSI feedback in FDD massive MIMO systems. By combining trained ternary quantization of attention projections with a learnable logarithmic activation (GLAQ), the proposed architecture substantially reduces UE-side parameter storage and computation while retaining and in some regimes improving reconstruction accuracy relative to full-precision attention baselines. Future work includes adaptive bitrate control, joint pilot, feedback optimization, and hardware co-design for ternary attention accelerators.

Acknowledgment

First, this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under. 6G·Cloud Research and Education Open Hub

(IITP-2025-RS-2024-00428780) and by the Korea government (MSIT) and Korea Institute for Advancement of Technology (KIAT). Second, the work funded by the Korea Government (Ministsry of Education) (P0025681-G02P22450002201-10054408, Semiconductor-Specialized University). Lastly, the work supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00404972, Development of 5G-A vRAN Research Platform).

REFERENCES

[1] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.

[2] Y. Cui, A. Guo, and C. Song, "Transnet: Full attention network for csi feedback in fdd massive mimo system," *IEEE Wireless Communications Letters*, vol. 11, no. 5, pp. 903–907, 2022.

[3] Z. Lu, J. Wang, and J. Song, "Binary neural network aided csi feedback in massive mimo system," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1305–1308, 2021.

[4] Z. Lu, X. Zhang, H. He, J. Wang, and J. Song, "Binarized aggregated network with quantization: Flexible deep learning deployment for csi feedback in massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5514–5525, 2022.

[5] F. Li and B. Liu, "Ternary weight networks," *CoRR*, vol. abs/1605.04711, 2016. [Online]. Available: http://arxiv.org/abs/1605.04711

[6] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=S1_pAu9xl