ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11462692

PSQ-PMC: A HARDWARE-FRIENDLY QUANTIZATION SCHEME FOR SPIKE-BASED NEURAL RADIANCE

*Ranxi Lin, Jiayi Li, Zhihong Meng, Pingqiang Zhou**

School of Information Science and Technology, ShanghaiTech University, Shanghai, China

ABSTRACT

Neural Radiance Fields (NeRF) has emerged as the core paradigm for 3D scene representation due to its remarkable high-fidelity 3D reconstruction capabilities. However, it is hindered by issues such as long training times and high memory consumption. PATA-INGP attempts to strike a balance between high rendering quality and low time cost by integrating Spiking Neural Networks (SNNs) and Instant-NGP NeRF (INGP-NeRF). Despite significantly reducing energy consumption during inference, the membrane potential of spike neurons occupies a large portion of memory resources. To address this, we propose a hardware-friendly quantization scheme that is tailored for the spike-based NeRF model. This scheme effectively alleviates the problems of gradient mismatch and large fluctuations in the output spikes. Experiments on various datasets demonstrate that our scheme far surpasses the other PTQ scheme and achieves considerable rendering quality compared to the QAT scheme, even under 4-bit widths.

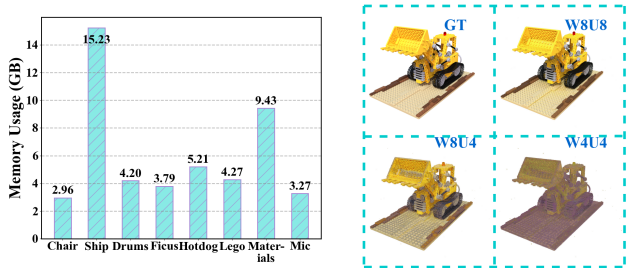
***Index Terms*—** Spiking Neural Network, Neural Rendering, Hardware-Friendly Quantization

1. INTRODUCTION

Neural Radiance Fields (NeRF) [1] implicitly represent 3D scenes using the coordinates of sample points and their viewing directions, allowing high-quality renderings and applications for various tasks. However, the complex network and intensive ray sampling consume a large number of computing and storage resources, limiting its application. To address this issue, TensoRF [2] uses low-rank tensor decomposition and adaptive sampling to accelerate volume rendering. Instant-NGP NeRF (INGP-NeRF) [3] employs multiresolution hash encoding to accelerate the generation of sample points, reducing redundant computations through coarse-to-fine spatial processing and feature fusion. However, the intensive multiply-accumulate (MAC) operations still resulted in significant energy consumption and storage pressure.

PATA-INGP [4] combines Spiking Neural Networks (SNNs) with INGP-NeRF to reduce energy consumption during inference, balancing rendering quality and inference time. SNNs utilize single-bit spikes for information transfer, reducing MAC operations and enhancing energy efficiency. However, the membrane potential of spike neurons leads to high memory usage. For example, in the Synthetic-NeRF dataset [1], Figure 1(a) shows the storage for the membrane potential of a single image, which far exceeds the model parameters (≈50MB).

To address the issue of high memory usage for membrane potential, the application of model compression techniques becomes highly essential. Quantization is a prevalent network compression method that mitigates memory consumption by compressing both model parameters and activation values into lower fixed-point numbers. Post-training quantization (PTQ) is more efficient than quantization-aware training (QAT) [5, 6] as it requires fewer parameter updates and epochs for quantization training [7]. Nevertheless, despite the fact that numerous studies have already achieved low-bit quantization in SNNs on classification tasks [8–12], the spike-based INGP-NeRF model's output of numerical values rather than spikes makes it vulnerable to performance degradation from quantization errors.



(a) Memory Usage (b) Qualitative results

Fig. 1. (a) Total memory usage of membrane potential for rendering a single image with a resolution of 800×800 under different scenes. (b) Qualitative results of Adaround under different bit widths.

Due to the compact MLPs in PATA-INGP, traditional block-wise quantization is inapplicable, so we adopt Adaround as the quantization baseline—using layer-wise training with a learnable vector $\nu \in \mathbb{R}^{M,K}$ for adaptive rounding during inference [13]. However, as shown in Figure 1(b), low-bit quantization causes obvious rendering distortion, attributed to two issues: 1) The quantizer uses the straight-through estimator (STE) for backpropagation, and the gradient is fixed at 1.0, which obviously deviates from the actual gradient and is exacerbated with the decrease of bit widths [14–16], 2) there is a significant difference between the output of the quantized model and that of the original model. Additionally, data range scaling introduces extra MAC operations during inference, conflicting with SNNs' characteristics.

To address the aforementioned challenges, achieving the qualified rendering quality and minimizing the storage usage, we first propose a **P**rogressive **S**tage-wise **Q**uantization (PSQ) method to alleviate the gradient mismatch problem and integrate it with global supervision to pursue the joint optimization of global and local aspects. After that, to mitigate the negative impact of quantization errors, we proposed the **P**ost-spike **M**embrane potential **C**lamp (PMC) strategy, which can significantly weaken the difference in output distribution between the quantized model and the original one without imposing any additional computational burden during inference. Hardware-friendly quantization is employed to eliminate the additional MAC

*Corresponding author.

operations during the quantization process [17]. Our contributions can be summarized as follows:

- We propose a novel PSQ method to alleviate gradient mismatch in low-bit quantization via staged optimization of linear layers and spike neurons, integrated with global supervision for joint local-global optimization;
- We propose a novel PMC strategy to reduce distribution differences between quantized and original spike output without extra inference cost, mitigating quantization errors impacts;
- The results on different datasets demonstrate the effectiveness of the proposed scheme. Even when compared with the QAT method, our approach can still achieve comparable results under a more aggressive compression ratio.

2. PRELIMINARIES

2.1. Spike-based Neural Radiance Field

The rendering pipeline of the original Instant-NGP NeRF mainly consists of three parts: a hash-coding multi-resolution grid, two compact MLPs, and the volume rendering process. The feature grid receives the indices x and interpolates the feature vectors Φ_θ to get the coordinate embedding $\mathbf{f}$. Then the two compact MLPs output the density σ and color c of each sample point:

$$\mathbf{f} = \text{interp}(x, \Phi_\theta) \tag{1}$$

$$h, \sigma = \text{MLP}_\sigma(\mathbf{f}), c = \text{MLP}_c(\mathbf{d}, h) \tag{2}$$

Where h means the intermediate feature, $\mathbf{d}$ denotes the feature of viewing direction after being encoded by the Spherical Harmonics (SH) encoder. For the PATA-INGP, the activation layers of the MLP are replaced by spiking neurons, and the final σ and c are obtained through multiple time steps. At each time step, the dynamics of the spiking layers are shown in the following formulas:

$$v^l(t) = \begin{cases} i(t), & t = 0 \\ (1 - \frac{1}{\tau})v^l(t-1) + \frac{1}{\tau}i(t), & t > 0 \end{cases} \tag{3}$$

$$s^l(t) = v^l(t) \geq V_{th} \tag{4}$$

$$v^l(t) = v^l(t) - s^l(t)V_{th} \tag{5}$$

where τ represents the decay factor. By introducing the hybrid input mode, PATA boosts the contribution of the early time step. After the completion of the above process, the RGB of each ray is calculated by the volume rendering method:

$$\alpha_i = 1 - exp(-\sigma_i\delta_i),\ T_i = \prod_{j=1}^{i-1}(1 - \alpha_j) \tag{6}$$

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i\alpha_i\mathbf{c}_i \tag{7}$$

Where T_i represents the transmittance and $\delta_i = t_{i+1} - t_i$ denotes the distance between adjacent sampled points.

2.2. Hardware-friendly Quantization

Current symmetrical uniform quantization algorithms tend to utilize the fake quantization node to match the computation flow of the hardware [18]:

$$x_q(x, s) = s * \text{Clamp}(\text{Round}(\frac{\text{x}}{\text{s}}), -2^{b-1}, 2^{b-1} - 1) \tag{8}$$

$$s = \frac{x_{max} - x_{min}}{2^b - 1} \tag{9}$$

where x and x_q represent the original data and quantized data, respectively, s is the quantization scale, and b means the bits. Hardware-friendly quantization utilizes bit-shift operation to replace the expensive MAC operation in equation (9) by forcing the scale $s = 2^n$, which is adapted to the characteristics of the PATA-INGP.

3. METHOD

In Section 3.1, we analyzed the quantization priorities of SNN layers and proposed the PSQ scheme to address the gradient mismatch issue. In Section 3.2, leveraging the single-bit output characteristic of SNNs, we proposed the PMC scheme to enhance the output matching between the quantized and original models. In Section 3.3, we designed the interpolation training scheme for discrete scaling factors and the scaling training for membrane potential thresholds.

3.1. Progressive Stage-wise Quantization

As shown in Figure 2, to address the gradient mismatch problem, we propose the progressive stage-wise quantization (PSQ) strategy and jointly optimize the model with the local reconstruction loss and global task loss. Given that the output of spike neurons is single-bit spikes, their input can be regarded as the result of superimposing the sampling from the prior linear layer's weights. Thus, the priority of reducing the quantization errors of weights is higher than that of membrane quantization. Regarding the local reconstruction loss, firstly, only the quantization nodes in the linear layer are activated. Thus, the gradients of the quantization parameters in the linear layer won't be disturbed by the STE in spike neurons. Next, we freeze the parameters in the linear layer and activate the quantization nodes in spike neurons, and only update the parameters in spike neurons until the end of the rest of the iterations. Updating the parameters in spike neurons alone enables them to learn and absorb the residual quantization perturbations.

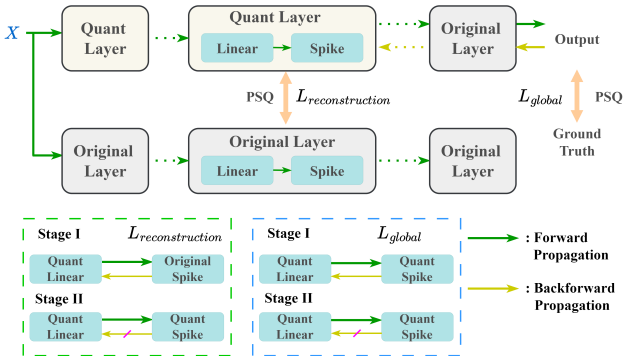


Fig. 2. Framework of the proposed PSQ scheme.

Global supervision plays a crucial role in ensuring that the optimization goals align with the actual ones [19–21]. We calculate the Mean-Square-Error (MSE) loss to quantify the discrepancy between the quantization model and the ground truth. Specifically, to achieve better synergy with the reconstruction loss, we activate the

quantization node within the linear layer and the spike neurons, aiming to evaluate the impact of quantization. During this process, the gradient of the global loss L_{global}, combined with that from the reconstruction loss, is utilized to update the quantization parameters. This collaborative update mechanism effectively prevents the model from being trapped in a local optimum. Subsequently, in the second stage, we keep freezing the parameters within the linear layer. This staged training strategy helps to improve the overall performance and stability of the model during the quantization process. To sum up, the loss function is defined as follows:

$$L_{quant} = \underbrace{||S_o^l - S_q^l||_{\mathrm{F}}^2 + f_{reg}(\nu)}_{L_{reconstruction}} + \underbrace{\mathrm{MSE}(O_q, gt)}_{L_{global}} \tag{10}$$

$$f_{reg}(\nu) = \sum_{i,j} 1 - |2h(\nu_{i,j}) - 1|^{\beta} \tag{11}$$

Where S_q^l and S_o^l denote the output spike of the l-th layer for the quantization model and original, respectively, $||\cdot||_{\mathrm{F}}$ is the Frobenius norm, O_q means the output of the quantization model, $h(\cdot)$ represents the rectified sigmoid function [22].

3.2. Post-spike Membrane Potential Clamp

We record the mean spike firing rate of the top 32 neurons in the first spike layer of MLP_σ, as depicted in Figure 3(a). We observe that

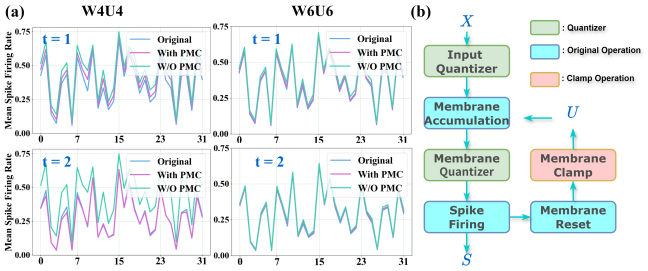


Fig. 3. (a) The output spike of the first 32 neurons in the first spiking layer in MLP_σ under different bit widths and time steps. (b) The quantized spike neurons computation flow.

both the increase in the time step and the decrease in the bit width contribute to the enlargement of quantization errors. This, in turn, leads to a continuous deterioration in the output differences between the quantization model and the original model. Considering that the spike firing is solely determined by whether the membrane potential exceeds the threshold, we note the following scenarios. If the magnitude of the membrane potential is sufficiently large or small, the neuron will either continue firing or remain inactive in the subsequent time step. Conversely, when the value of the membrane potential falls within a specific range, the activity of the neuron will be jointly decided by the current membrane potential and the incoming input in the next time step. Based on these insights, as shown in Figure 3(b), we propose the Post-spike Membrane potential Clamp (PMC) strategy. For the l-th spiking layer with K neurons, we set:

$$v^l(t) = v^l(t).clamp(-V_c * \mathrm{Sig}(\alpha_{min}^l), V_c * \mathrm{Sig}(\alpha_{max}^l)) \tag{12}$$

Where $v^l(t)$ denotes the quantized membrane potential after spike firing and reset, V_c is the boundary of the membrane potential and is set to 4.0 during the training, $\alpha_{min}^l \in \mathbb{R}^K$ and $\alpha_{max}^l \in \mathbb{R}^K$ represents the learnable parameters, and sig means the sigmoid function.

Table 1. Quantization results under different bit widths on different datasets. W/U represents the bit widths of parameters in the model and membrane potential.

Synthetic-NeRF			Mip-NeRF 360		
W/U	PSNR↑	SSIM↑	W/U	PSNR↑	SSIM↑
32/32	31.78	0.954	32/32	24.47	0.631
8/8	30.95	0.949	8/8	24.30	0.629
6/6	29.71	0.945	7/7	24.19	0.626
8/4	28.82	0.938	8/6	24.25	0.627
4/4	24.75	0.904	6/6	23.22	0.605

PMC limits the range of the membrane potential after spike firing and reset. It doesn't interfere with spike generation and directly removes extreme values that have no impact on spike firing activity in the next time step. In Figure 3(a), after applying the proposed PMC strategy, the distance of the mean value of spike firing rate between the quantized model and the original model is significantly reduced. Note that PMC is only applied during training and won't alter the computation graph of spike neurons in inference. As PMC effectively alleviates the disturbance from quantization errors, it allows the model to focus more on the membrane potential range near the threshold.

3.3. Optimization of Scaling Factor and Threshold

All scaling factors are initialized by minimizing the L2 distance between the quantized tensor and the original one. Since the value of the scaling factor $s = 2^n$, we directly optimize the power item $n \in \mathbb{R}^1$ through interpolation:

$$n_f = \mathrm{Floor}(n), n_c = \mathrm{Ceil}(n) \tag{13}$$

$$x_{q,f} = x_q(x, 2^{n_f}), x_{q,c} = x_q(x, 2^{n_c}) \tag{14}$$

$$x_q = (1 - n + n_f) * x_{q,f} + (n - n_f) * x_{q,c} \tag{15}$$

During inference we set $s = 2^{\mathrm{round}(n)}$. We only set the scaling factor in MLPs as trainable parameters, and each quantizer uses tensor-wise quantization to promote parallelism and avoid additional storage burden.

Quantization errors lead to changes in the membrane potential distribution. Besides the PMC mechanism, we have set learnable parameters to scale the threshold:

$$V_{thq} = \mathrm{Sig}(\eta) * V_{th} \tag{16}$$

Where V_{th} is the threshold of spike neurons, $\eta \in \mathbb{R}^1$ is the learnable parameter and is set to 2.5 as the initial value. The threshold adopts the same scaling factor as the membrane potential.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

Models & Metrics & Datasets: We conduct the experiments based on PATA-INGP with the average time step of 4 and evaluate our method on two established benchmark datasets: the Synthetic-NeRF dataset [1] and the Mip-NeRF 360 dataset [23]. We use both Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) as quantitative metrics.

Training setting: We configure each MLP with a hidden layer size of 64 and derive the original model by adhering to the settings

of PATA-INGP[1]. In quantization training, each layer undergoes 4000 iterations with a fixed learning rate of 1×10^{-3}, using the Adam optimizer to update quantization parameters. For α_{max} and α_{min} in Equation 12, they are initialized as zero vectors. Feature grids use the same quantization bit width as weights, with each level quantized independently. Symmetric uniform quantizers are applied to feature grids, weights, membrane potentials, and sigmoid/exp functions. Decay factors are directly mapped to the nearest power-of-two values. Model parameters remain frozen during training.

4.2. Results

We present the performance under different experimental sets in Table 2. Since the memory usage of the membrane potential is far higher than the parameters in the model, we also report the result of mixed-precision quantization on different datasets. When using 8-bit quantization, the decrease in PSNR is only 0.83, while the rendering quality on the Mip-NeRF 360 dataset remains almost unchanged. Even with 4-bit quantization, our strategy can still maintain a certain level of rendering capability. For the mixed-precision quantization, compared to the original model, the case of W8U4 can achieve a 87.5% drop in memory overhead of inference and more than a 75% drop in parameter storage of the model with the only cost of a 9.3% loss in PSNR, highlighting the excellent inference efficiency.

Table 2. Quantization results under different bit widths on different datasets; the numbers in parentheses represent the drop in PSNR compared to the original model. For the schemes based on QAT, we mark them in blue, and the best results are highlighted in bold.

Method	Synthetic-NeRF		Mip-NeRF 360	
	FQR↓	PSNR↑	FQR↓	PSNR↑
LSQ+ [5]	9.60	**32.11**(**−0.31**)	9.60	**25.48**(**−0.07**)
A-CAQ [5]	**7.76**	32.00 (-0.42)	**7.11**	25.40(-0.15)
PTQ [24]	9.60	31.98 (-0.46)	9.60	25.38(-0.17)
Ours	8.00	30.95 (-0.83)	8.00	24.30(-0.17)
LSQ+	6.60	25.06 (-7.36)	7.60	23.84(-1.71)
A-CAQ	**5.33**	27.58 (-4.84)	**6.86**	25.18(-0.27)
PTQ	6.60	22.29 (-10.13)	7.60	22.62(-2.93)
Ours	6.00	**29.71**(**−2.07**)	7.00	**24.25**(**−0.22**))
LSQ+	5.60	20.26 (-12.16)	6.60	22.44(-3.11)
A-CAQ	4.74	**25.96**(**−6.46**)	6.58	**24.74**(**−0.81**)
PTQ	5.60	17.75 (-14.67)	6.60	17.52(-8.03)
Ours	**4.00**	24.75 (-7.03)	**6.00**	23.22(-1.08)

Due to the absence of any other quantization schemes in the experiments conducted on the spike-based NeRF model, we directly compare the performance with other quantization methods on the original INGP-NeRF model, including PTQ and QAT methods, as shown in Table 2. The feature quantization rate (FQR) [5] is introduced to measure the performance. In the configuration of high bit widths, our strategy can achieve a comparable PSNR with other quantization methods. As the bit width decreases, the advantages of our scheme become more pronounced and even surpass those of the QAT scheme. More importantly, our scheme not only employs tensor-wise quantization in all quantizers to minimize the storage load of quantization parameters but also avoids the MAC operations of quantization nodes during inference, resulting in lower power consumption compared to other schemes, which is perfectly in line with the characteristics of SNNs.

[1]Code: https://github.com/ToumaKazusa2805/QuanPATA

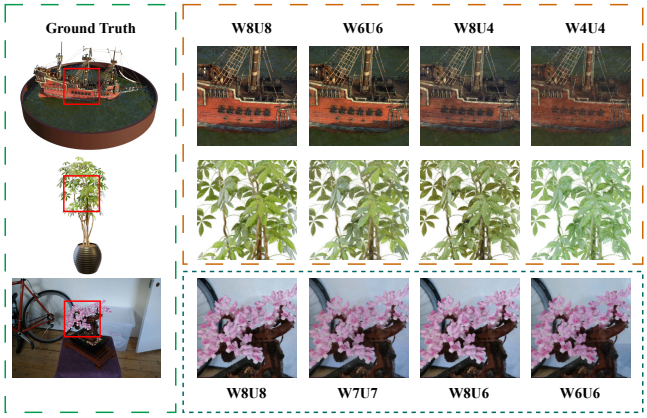


Fig. 4. Qualitative results for the ship, ficus, and bonsai scenes under different bit widths.

Figure 4 shows the qualitative results under different bit widths. It can be seen that even though the membrane potentials are quantized to 4 bits, the render result is still clean and has no obvious structural distortion.

4.3. Ablation Study

In Table 2, the configuration of W8U4 strikes the best balance between rendering quality and memory compression rate in comparison to others. Hence, we select it as the baseline for the ablation experiment. As shown in Table 3, the original Adaround algorithm gets the worst quality compared to our scheme. Since PMC can effectively eliminate the negative impact of quantization errors without incurring an additional computational burden, even when used alone, it can still achieve significant improvements. The proposed PSQ training scheme has also demonstrated its effectiveness, highlighting the significance of gradient correction.

Table 3. Ablation study on the Synthetic-NeRF Dataset under W8U4.

PSQ	PMC	PSNR↑	SSIM↑
×	×	24.56	0.881
✓	×	25.16	0.908
×	✓	28.25	0.934
✓	✓	28.82	0.938

5. CONCLUSION

In this paper, we present a hardware-friendly quantization scheme for spike-based NeRF models. First, we propose the PSQ strategy to address the gradient mismatch caused by the STE, optimizing quantization parameters via the collaborative optimization of global loss and local reconstruction loss. Second, the PMC strategy effectively mitigates membrane potential perturbations induced by quantization errors, thereby reducing discrepancies between the quantized and original outputs. Experimental results on different datasets demonstrate that our scheme can effectively alleviate the storage pressure of model parameters and reduce runtime memory consumption.

6. ACKNOWLEDGMENT

This work was supported by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 24JD1402500.

7. REFERENCES

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su, "Tensorf: Tensorial radiance fields," in *European conference on computer vision*. Springer, 2022, pp. 333–350.

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[4] Ranxi Lin, Canming Yao, Jiayi Li, Weihang Liu, Xin Lou, and Pingqiang Zhou, "Adaptive time-step training for enhancing spike-based neural radiance fields," *arXiv preprint arXiv:2507.23033*, 2025.

[5] Weihang Liu, Xue Xian Zheng, Jingyi Yu, and Xin Lou, "Content-aware radiance fields: Aligning model complexity with scene intricacy through learned bitwidth quantization," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–256.

[6] Weihang Liu, Xue Xian Zheng, Yuke Li, Tareq Y Al-Naffouri, Jingyi Yu, and Xin Lou, "Coarf++: Content-aware radiance field aligning model complexity with scene intricacy," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[7] Ahmed Hasssan, Anupreetham Anupreetham, Jian Meng, and Jae-sun Seo, "Quant-nerf: Efficient end-to-end quantization of neural radiance fields with low-precision 3d gaussian representation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] Haomin Li, Fangxin Liu, Zewen Sun, Zongwu Wang, Shiyuan Huang, Ning Yang, and Li Jiang, "Neuronquant: Accurate and efficient post-training quantization for spiking neural networks," in *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, 2025, pp. 734–740.

[9] Ruokai Yin, Yuhang Li, Abhishek Moitra, and Priyadarshini Panda, "Mint: Multiplier-less integer quantization for energy efficient spiking neural networks," in *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2024, pp. 830–835.

[10] Dehao Zhang, Shuai Wang, Yichen Xiao, Wenjie Wei, Yimeng Shan, Malu Zhang, and Yang Yang, "Memory-free and parallel computation for quantized spiking neural networks," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[11] Hila Naaman, Neil Irwin Bernardo, Alejandro Cohen, and Yonina C Eldar, "Time encoding quantization of bandlimited and finite-rate-of-innovation signals," *IEEE Access*, 2025.

[12] Dorian Florescu, "Model-driven quantization for time encoding machines," in *2023 International Conference on Sampling Theory and Applications (SampTA)*. IEEE, 2023, pp. 1–5.

[13] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort, "Up or down? adaptive rounding for post-training quantization," in *International conference on machine learning*. PMLR, 2020, pp. 7197–7206.

[14] Yoshua Bengio, Nicholas Léonard, and Aaron Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[15] Yanjing Li, Sheng Xu, Mingbao Lin, Xianbin Cao, Chuanjian Liu, Xiao Sun, and Baochang Zhang, "Bi-vit: Pushing the limit of vision transformer quantization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 3243–3251.

[16] Jixing Li, Gang Chen, Min Jin, Wenyu Mao, and Huaxiang Lu, "Ae-qdrop: towards accurate and efficient low-bit post-training quantization for a convolutional neural network," *Electronics*, vol. 13, no. 3, pp. 644, 2024.

[17] Sung-En Chang, Yanyu Li, Mengshu Sun, Runbin Shi, Hayden K-H So, Xuehai Qian, Yanzhi Wang, and Xue Lin, "Mix and match: A novel fpga-centric deep neural network quantization framework," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 208–220.

[18] Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori, "A comprehensive survey on model quantization for deep neural networks in image classification," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 6, pp. 1–50, 2023.

[19] Sicheng Li, Hao Li, Yiyi Liao, and Lu Yu, "Nerfcodec: Neural feature compression meets neural radiance fields for memory-efficient scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21274–21283.

[20] Tuan Pham and Stephan Mandt, "Neural nerf compression," in *International Conference on Machine Learning*. PMLR, 2024, pp. 40592–40610.

[21] Jiawei Liu, Lin Niu, Zhihang Yuan, Dawei Yang, Xinggang Wang, and Wenyu Liu, "Pd-quant: Post-training quantization based on prediction difference metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24427–24437.

[22] Christos Louizos, Max Welling, and Diederik P Kingma, "Learning sparse neural networks through l_0 regularization," *arXiv preprint arXiv:1712.01312*, 2017.

[23] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[24] Chaolin Rao, Huangjie Yu, Haochuan Wan, Jindong Zhou, Yueyang Zheng, Minye Wu, Yu Ma, Anpei Chen, Binzhe Yuan, Pingqiang Zhou, et al., "Icarus: A specialized architecture for neural radiance fields rendering," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–14, 2022.