

Quantum Surrogate-Driven Image Classifier: A Gradient-Free Approach to Avoid Barren Plateaus

Yichen Xie
La Salle College, HKSAR
s21325@lsc.hk

Abstract—Training deep quantum neural networks (QNNs) for image classification is notoriously difficult due to vanishing gradients (barren plateaus) and limited nonlinearity in purely unitary circuits. We propose a novel gradient-free surrogate-driven framework combined with mid-circuit measurement and reset of ancillary qubits to induce effective nonunitarity. Our approach uses a classical neural surrogate to predict measurement outcomes from circuit parameters to avoid direct gradients. Theoretical results prove that bypassing quantum gradients mitigates plateau issues. Experiments on MNIST, CIFAR-10, and CIFAR-100 with 15-qubit, 6-layer circuits using four resettable ancillas demonstrate superior accuracy compared to direct-gradient QNNs and classical baselines. Our method also serves as a potential for a generalized training framework applicable to various QNN architectures beyond image classification.

I. INTRODUCTION

Parametric quantum circuits (PQCs) leverage the exponential dimensionality of an n -qubit Hilbert space, offering powerful potential for tasks such as image classification. However, their practical training faces critical challenges: the barren plateau phenomenon, causing gradients to vanish exponentially with increasing circuit depth or qubit count, severely impeding gradient-based optimization [1], [2]; and the inherent linearity of unitary operations within PQCs, restricting their expressibility compared to classical deep networks that utilize nonlinear activations [3], [4].

To address these challenges, we propose a novel surrogate-based optimization approach. Instead of directly computing quantum gradients, we train a classical neural surrogate model to learn the mapping from circuit parameters to measurement outcomes, thus providing efficient surrogate gradients without quantum differentiation [5], [6]. Furthermore, we introduce mid-circuit measurement and reset operations on designated ancilla qubits after each circuit layer, creating branching effects analogous to nonlinear activations in classical neural networks, and facilitating resource-efficient qubit reuse [7], [8].

Additionally, recognizing that naive amplitude encoding is inefficient for large images, we first compress input images into a reduced feature space (e.g., 256 for MNIST, 512 for CIFAR), then use a parameter-projection network to map these compressed features into an extensive set of circuit parameters [9]–[11]. Our integrated approach, combining surrogate-driven optimization, nonlinear-like mid-circuit operations, and efficient data encoding, demonstrates superior performance compared to direct-gradient quantum neural networks and classical baselines, as validated by extensive simulations on

benchmark datasets including MNIST, CIFAR-10, and CIFAR-100 [4], [6], [11].

II. ARCHITECTURE AND THEORETICAL FRAMEWORK

A. Overview

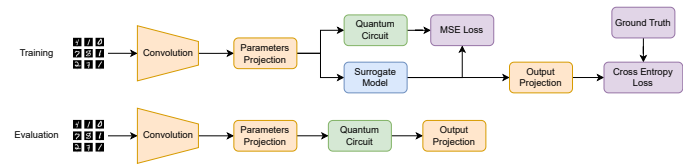
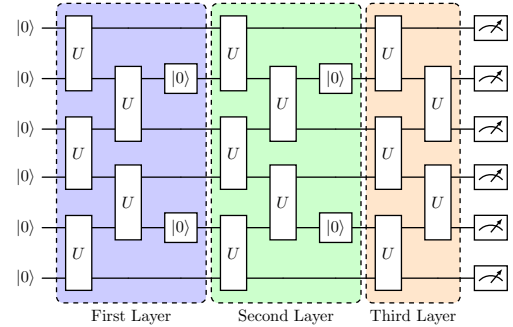


Fig. 1: Framework of the architecture during training and evaluation.

We now describe the overall architecture. The quantum circuit has $n = 15$ qubits, including $n_a = 4$ ancillas at wires 3,6,9,12, across $L = 6$ layers. Each layer applies parameterized unitaries, measures ancillas, resets them to $|0\rangle$ if measured as $|1\rangle$, and proceeds to the next layer. A simplified 6-qubit, 3-layer version is shown in Figure 2.



where

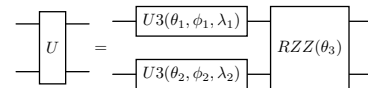


Fig. 2: Simplified quantum circuit with 6 qubits, 3 layers, and ancillas at wires 1,5. Each U module has 7 parameters.

For MNIST ($28 \times 28 = 784$ pixels), images are compressed to 256 features; for CIFAR-10/100 ($32 \times 32 \times 3 = 3072$), they are compressed to 512 features, using a simple VGG-style CNN $C_w(x)$ consisting of four blocks with Batch Normalization, ReLU activations [12], [13], and max pooling. The

compressed vector $\mathbf{z} \in \mathbb{R}^{256}$ or \mathbb{R}^{512} is then fed into a two-layered MLP $\Gamma_w(\mathbf{z}) \in \mathbb{R}^p$, generating p trainable angles θ (with $p = 588$ for 6 layers and 15 qubits). By measuring ancillas after each layer and resetting them, the main qubits undergo an effective nonlinear transformation (Theorem 1).

The circuit output (expectation value with Pauli-Z observable) is passed into a projection layer $O_w(x) \in [0, 1]$ for classification. We compute the cross-entropy (CE) loss:

$$\mathcal{L}_1(x, y) = - \sum_{j=1}^C y_j \log(O_w(S_w(\Gamma_w(C_w(x_j)))))) \quad (1)$$

where S_w is the surrogate model, x is the input, y_j is the one-hot label, and C is the number of classes. Gradients ∇O_w are obtained by backpropagating $\nabla \mathcal{L}_1$.

To avoid gradients through the quantum circuit during training, we sample parameter points around θ_t , run the circuit, and fit a surrogate neural network $S_w(\theta) \approx \mathbf{m}(\theta)$ using mean-squared error (MSE):

$$\mathcal{L}_2(\theta) = \frac{1}{2p} \sum_{i=1}^p (S_w(\theta_i) - \mathbf{m}(\theta_i))^2. \quad (2)$$

This surrogate provides differentiable approximations of quantum outcomes, allowing efficient classical updates to CNN and MLP parameters.

Traditional QNNs apply purely unitary layers, measuring only at the end, resulting in linear transformations and susceptibility to barren plateaus [1]. Our method introduces mid-layer ancilla measurements, inducing nonunitary transformations, and circumvents barren plateaus by using gradient-free surrogate modeling.

B. Nonunitary Gates by Measuring and Resetting Ancillas

We first prove that measuring and resetting a subset of qubits can implement a nontrivial class of nonlinear maps on the remaining subsystem without collapsing the entire system to a single state.

Lemma 1 (Non-unitary but Non-collapsing Layer). *Let $n = n_m + n_a$ qubits be divided into n_m main qubits and n_a ancillas ($n_a < n$). Start from an n -qubit density operator ρ . Apply a global unitary $U(\theta)$, then measure every ancilla in the $\{|0\rangle, |1\rangle\}$ basis. For each outcome $x \in \{0, 1\}^{n_a}$ reset the ancillas to $|0^{\otimes n_a}\rangle$. Denote the overall completely-positive trace-preserving (CPTP) map by \mathcal{M} . Tracing out the ancillas yields the channel*

$$\text{Tr}_{\text{anc}}[\mathcal{M}(\rho)] = \sum_{x \in \{0, 1\}^{n_a}} K_x \rho K_x^\dagger, \quad (3)$$

where each Kraus operator on the main subsystem is

$$K_x = {}_{\text{anc}}\langle 0^{\otimes n_a} | (|0^{\otimes n_a}\rangle\langle x| \otimes I_{n_m}) U = {}_{\text{anc}}\langle x | U. \quad (4)$$

Because the K_x are generically non-unitary and (for entangling U) mutually distinct, \mathcal{M} performs a nonlinear, non-collapsing transformation on the main qubits that cannot be written as a single unitary acting only on those qubits.

Proof. Let $M_x = (|x\rangle\langle x|)_{\text{anc}} \otimes I_{n_m}$ project onto the ancilla outcome x . Define the *reset operator*

$$R_x = (|0^{\otimes n_a}\rangle\langle x|)_{\text{anc}} \otimes I_{n_m}, \quad x \in \{0, 1\}^{n_a}. \quad (5)$$

For each branch x we evolve as $\rho \mapsto R_x M_x U \rho U^\dagger M_x R_x^\dagger$. Summing over outcomes gives the global CPTP map

$$\mathcal{M}(\rho) = \sum_x R_x M_x U \rho U^\dagger M_x R_x^\dagger. \quad (6)$$

Tracing out the ancillas and inserting the resolutions of the identity,

$$\begin{aligned} \text{Tr}_{\text{anc}}[\mathcal{M}(\rho)] &= \sum_x \text{Tr}_{\text{anc}}[(R_x M_x U) \rho (U^\dagger M_x R_x^\dagger)] \\ &= \sum_x ({}_{\text{anc}}\langle 0^{\otimes n_a} | R_x M_x U) \rho (U^\dagger M_x R_x^\dagger | 0^{\otimes n_a}\rangle_{\text{anc}}) \\ &= \sum_x K_x \rho K_x^\dagger, \end{aligned} \quad (7)$$

with K_x defined in (4). The second equality uses the fact that $(|y\rangle\langle y|)_{\text{anc}}$ appearing in the partial trace selects $y = 0^{\otimes n_a}$ because R_x resets ancillas to that state.

Unitarity of U and orthogonality of the projectors imply $\sum_x K_x^\dagger K_x = I_{n_m}$, so the set $\{K_x\}_{x \in \{0, 1\}^{n_a}}$ constitutes a valid Kraus representation of a CPTP map on the main qubits.

If U entangles ancillas with main qubits, the operators K_x differ for different x and are not proportional to unitaries on the main subsystem. Multiple Kraus terms therefore survive in the sum, demonstrating that \mathcal{M} is generally a non-unitary, non-projective transformation that nonetheless leaves the main qubits in a mixed state retaining information from all measurement branches. \square

Corollary 1 (Ancilla reuse after measure–reset). *Using notation from Lemma 1, the full post-measurement-reset state factorizes as*

$$\mathcal{M}(\rho) = |0^{\otimes n_a}\rangle\langle 0^{\otimes n_a}|_{\text{anc}} \otimes \sum_{x \in \{0, 1\}^{n_a}} K_x \rho K_x^\dagger. \quad (8)$$

Thus, the ancillas are deterministically reset to $|0^{\otimes n_a}\rangle$, remain unentangled from the main register, and can be safely reused.

Proof. Start from $\mathcal{M}(\rho) = \sum_x (R_x M_x U) \rho (U^\dagger M_x R_x^\dagger)$, substitute $R_x M_x = (|0^{\otimes n_a}\rangle\langle x|) \otimes I_{n_m}$, and insert identity resolution $I_{\text{anc}} = \sum_y |y\rangle\langle y|$. Only $y = 0^{\otimes n_a}$ remains, yielding the factorization with $K_x = {}_{\text{anc}}\langle x | U$. \square

Theorem 1 (Nonlinearity and Preserved Expressibility). *The measure-and-reset operation from Lemma 1 is inserted after each layer in a deep circuit of total depth L . Let $\Phi(\theta)$ be the resulting overall map from initial n_m -qubit states to final n_m -qubit states. Then for sufficiently large $n_a \geq 1$ and entangling unitaries $U_\ell(\theta_\ell)$, the map Φ can realize a broad class of nonlinear transformations on the main qubits, while avoiding*

deterministic collapse onto a single pure state. Formally, there exist Kraus representations:

$$\text{Tr}_{\text{anc}} \left[\prod_{\ell=1}^L \left(R \circ \mathcal{M}_\ell \circ U_\ell(\theta_\ell) \right) \right] = \sum_k V_k(\cdot) V_k^\dagger \quad (9)$$

with V_k nontrivial, and no single Kraus operator dominates all inputs. Hence, this multi-layer measure-and-reset design acts as an effectively nonlinear feedforward QNN, which can approximate a wide range of channel mappings without collapsing distinct inputs into a single final state.

Proof. We recursively apply Lemma 1. Each layer ℓ has a global unitary U_ℓ on n qubits, followed by measuring ancillas plus reset. The partial trace over ancillas after L layers yields $\sum_k V_k(\cdot) V_k^\dagger$ with V_k formed by branching sequences of M_x projectors. If the unitaries $U_\ell(\theta_\ell)$ sufficiently entangle ancillas with main qubits, measurement outcomes condition distinct Kraus branches. Because resets restore the ancillas to $|0 \cdots 0\rangle$, each layer reintroduces nonlinearity. Non-collapse follows from the fact that no single outcome branch has unit probability for all states (unless gates are trivial). Thus multiple Kraus operators remain active, preserving variation among inputs. \square

C. Surrogate Model for Gradient-Free Parameter Updates

Next, we outline a gradient-free procedure to train these multi-layer measure-and-reset circuits. Denote the circuit parameters by θ , and let the final measured probabilities on n_m qubits for a classical label y be $p(y|\theta; x)$, where x is the input data. In classification, we aim to minimize a cost function:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{(x_i, y_i)} \ell(p(\cdot|\theta; x_i), y_i), \quad (10)$$

which is a cross-entropy in our case. A typical gradient-based approach would compute $\partial \mathcal{L} / \partial \theta_j$ via parameter-shift rules, but for large n or deep L , these gradients can vanish due to barren plateaus [1].

Instead, we train a classical neural surrogate $S(\theta)$ to predict the measurement outcomes of the quantum circuit (not the final loss). Specifically, let the circuit output a vector $\mathbf{m}(\theta, x) \in \mathbb{R}^d$ of measurement statistics (e.g. log-odds for each class), and let S be a neural network that takes (θ, x) as input and estimates $\mathbf{m}(\theta, x)$. We train S by sampling θ in a neighborhood of the current parameter and measuring the quantum device on data x to get true outcomes \mathbf{m} . Then S is optimized (in a purely classical sense) to fit these measured samples:

$$\min_w \sum_k \|\mathbf{m}(\theta^{(k)}, x^{(k)}) - S_w(\theta^{(k)}, x^{(k)})\|^2. \quad (11)$$

Once S accurately approximates \mathbf{m} near the current θ , we compute the classification loss purely classically:

$$\tilde{\mathcal{L}}(\theta) = \frac{1}{N} \sum_{(x_i, y_i)} \ell(S(\theta, x_i), y_i), \quad (12)$$

and we take classical gradients of $\tilde{\mathcal{L}}$ w.r.t. θ . This bypasses direct quantum gradients. The optimization update is:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \tilde{\mathcal{L}}(\theta) \big|_{\theta=\theta_t}. \quad (13)$$

Because the neural surrogate can avoid the exponential gradient decay typical in quantum circuits, we do not encounter vanishing gradients from the circuit itself. We will see in experiments that this method yields stable updates even for deeper circuits.

Below is a lemma showing that if S is sufficiently expressive and well-trained, then descending on $\tilde{\mathcal{L}}$ descends on the true \mathcal{L} as well:

Lemma 2 (Surrogate Descent). *Assume the true loss $L(\theta)$ is continuously differentiable. Let $S(\theta)$ be a surrogate model such that $S(\theta^{(t)}) = L(\theta^{(t)})$ and $\|\nabla_\theta S(\theta) - \nabla_\theta L(\theta)\| \leq \epsilon$ for all θ in a convex neighborhood \mathcal{N} around $\theta^{(t)}$. Then for a sufficiently small learning rate $\eta > 0$, the update $\theta^{(t+1)} = \theta^{(t)} - \eta \nabla S(\theta^{(t)})$ yields*

$$L(\theta^{(t+1)}) \leq L(\theta^{(t)}) - \frac{\eta}{2} \|\nabla L(\theta^{(t)})\|^2 + O(\eta^2), \quad (14)$$

provided $\theta^{(t+1)}$ stays in \mathcal{N} and ϵ is sufficiently small.

Proof. Since L is differentiable, a first-order Taylor expansion around $\theta^{(t)}$ gives:

$$L(\theta^{(t+1)}) = L(\theta^{(t)}) + \nabla L(\theta^{(t)})^\top (\theta^{(t+1)} - \theta^{(t)}) + R, \quad (15)$$

where R is the second-order remainder term: $R = \frac{1}{2} (\theta^{(t+1)} - \theta^{(t)})^\top H_L (\theta^{(t+1)} - \theta^{(t)})$ for some Hessian $H_L = \int_0^1 \nabla^2 L(\theta^{(t)} + \tau(\theta^{(t+1)} - \theta^{(t)})) d\tau$. There exists $M > 0$ such that $\|H_L\| \leq M$ if \mathcal{N} is bounded and L is twice continuously differentiable (by the extreme value theorem). Substituting the update $\theta^{(t+1)} = \theta^{(t)} - \eta \nabla S(\theta^{(t)})$ into (15), we get:

$$\begin{aligned} L(\theta^{(t+1)}) &= L(\theta^{(t)}) - \eta \nabla L(\theta^{(t)})^\top \nabla S(\theta^{(t)}) \\ &\quad + \frac{\eta^2}{2} \nabla S(\theta^{(t)})^\top H_L \nabla S(\theta^{(t)}) \\ &\leq L(\theta^{(t)}) - \eta \nabla L(\theta^{(t)})^\top \nabla S(\theta^{(t)}) \\ &\quad + \frac{\eta^2 M}{2} \|\nabla S(\theta^{(t)})\|^2. \end{aligned} \quad (16)$$

Using the assumption on gradient closeness: $\nabla S(\theta^{(t)}) = \nabla L(\theta^{(t)}) + \mathbf{e}$ with $\|\mathbf{e}\| \leq \epsilon$. Then

$$\begin{aligned} \nabla L(\theta^{(t)})^\top \nabla S(\theta^{(t)}) &= \|\nabla L(\theta^{(t)})\|^2 + \nabla L(\theta^{(t)})^\top \mathbf{e} \\ &\geq \|\nabla L(\theta^{(t)})\|^2 - \|\nabla L(\theta^{(t)})\| \|\mathbf{e}\| \\ &\geq \|\nabla L(\theta^{(t)})\|^2 - \epsilon \|\nabla L(\theta^{(t)})\|. \end{aligned}$$

Similarly, $\|\nabla S(\theta^{(t)})\| \leq \|\nabla L(\theta^{(t)})\| + \epsilon$. Inserting these into (16) yields:

$$\begin{aligned} L(\theta^{(t+1)}) &\leq L(\theta^{(t)}) - \eta \left(\|\nabla L(\theta^{(t)})\|^2 - \epsilon \|\nabla L(\theta^{(t)})\| \right) \\ &\quad + \frac{\eta^2 M}{2} (\|\nabla L(\theta^{(t)})\| + \epsilon)^2. \end{aligned} \quad (17)$$

For small enough η , the η^2 term is negligible compared to η (specifically, if $\eta < \frac{\|\nabla L(\boldsymbol{\theta}^{(t)})\|}{M}$, the second-order term is $O(\eta^2 \|\nabla L\|^2)$). Ignoring $O(\eta^2)$ and using ϵ small, we get approximately:

$$L(\boldsymbol{\theta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t)}) - \eta \|\nabla L(\boldsymbol{\theta}^{(t)})\|^2 + \eta \epsilon \|\nabla L(\boldsymbol{\theta}^{(t)})\|. \quad (18)$$

If we choose ϵ such that $\epsilon \ll \|\nabla L(\boldsymbol{\theta}^{(t)})\|$ in the current neighborhood (or more strictly, treat ϵ as $o(\|\nabla L\|)$), the last term becomes second-order small compared to the first term. Thus:

$$L(\boldsymbol{\theta}^{(t+1)}) < L(\boldsymbol{\theta}^{(t)}) - \frac{\eta}{2} \|\nabla L(\boldsymbol{\theta}^{(t)})\|^2, \quad (19)$$

for sufficiently small η and ϵ , completing the proof of a guaranteed decrease in L up to first order. \square

Lemma 2 implies that as long as our surrogate S captures the local slope of L with reasonable accuracy (small ϵ), moving opposite to ∇S will decrease the true loss L . In particular, even if ∇L is very small (barren plateau regime), ∇S might not be, because the surrogate can fit through sparse samples that detect slight differences in L . The condition $\boldsymbol{\theta}^{(t+1)}$ remains in \mathcal{N} is enforceable by taking sufficiently small steps or by using a trust region radius to limit how far we go based on the surrogate’s validity.

D. Convergence of Surrogate Optimization

Next, we address the convergence of the iterative surrogate-based training procedure. We will show that under the assumption that the surrogate model can be made arbitrarily accurate (by taking more sample points or a more flexible functional form as needed), the parameter sequence $\{\boldsymbol{\theta}^{(t)}\}$ produced by our method will approach a stationary point (usually a local minimum) of the true loss L .

Theorem 2 (Convergence). *Consider the iterative update $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \nabla S^{(t)}(\boldsymbol{\theta}^{(t)})$ where $S^{(t)}$ is a surrogate model fit around $\boldsymbol{\theta}^{(t)}$ as described above. Assume: 1) The loss $L(\boldsymbol{\theta})$ is lower bounded (i.e., has a global minimum L_{\min}) and L is L -Lipschitz smooth (its gradient is Lipschitz continuous with constant L); 2) At each iteration t , the surrogate $S^{(t)}$ satisfies $\|\nabla S^{(t)}(\boldsymbol{\theta}^{(t)}) - \nabla L(\boldsymbol{\theta}^{(t)})\| \leq \epsilon_t$, where ϵ_t can be made arbitrarily small by using a sufficiently accurate surrogate (and assume ϵ_t is indeed chosen small enough at each step); 3) The step size η_t is chosen via a diminishing schedule or sufficiently small constant such that $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$ (a typical condition for gradient descent convergence). Then $\{\boldsymbol{\theta}^{(t)}\}$ converges to a stationary point of L , i.e., $\lim_{t \rightarrow \infty} \nabla L(\boldsymbol{\theta}^{(t)}) = \mathbf{0}$, and $\lim_{t \rightarrow \infty} L(\boldsymbol{\theta}^{(t)}) = L^*$ for some local minimum value $L^* \geq L_{\min}$.*

Proof. Under the Lipschitz smoothness assumption for ∇L , we have the standard descent lemma for the true loss:

$$L(\boldsymbol{\theta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t)}) + \nabla L(\boldsymbol{\theta}^{(t)})^\top (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + \frac{L}{2} \|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|^2. \quad (20)$$

Substitute the update $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \nabla S^{(t)}(\boldsymbol{\theta}^{(t)})$:

$$\begin{aligned} L(\boldsymbol{\theta}^{(t+1)}) &\leq L(\boldsymbol{\theta}^{(t)}) - \eta_t \nabla L(\boldsymbol{\theta}^{(t)})^\top \nabla S^{(t)}(\boldsymbol{\theta}^{(t)}) \\ &\quad + \frac{L\eta_t^2}{2} \|\nabla S^{(t)}(\boldsymbol{\theta}^{(t)})\|^2 \\ &\leq L(\boldsymbol{\theta}^{(t)}) - \eta_t \|\nabla L(\boldsymbol{\theta}^{(t)})\|^2 \\ &\quad + \eta_t \|\nabla L(\boldsymbol{\theta}^{(t)})\| \|\mathbf{e}_t\| \\ &\quad + \frac{L\eta_t^2}{2} (\|\nabla L(\boldsymbol{\theta}^{(t)})\| + \|\mathbf{e}_t\|)^2, \end{aligned} \quad (21)$$

where $\mathbf{e}_t = \nabla S^{(t)}(\boldsymbol{\theta}^{(t)}) - \nabla L(\boldsymbol{\theta}^{(t)})$ and $\|\mathbf{e}_t\| \leq \epsilon_t$. If ϵ_t is made sufficiently small at each iteration (ideally $\epsilon_t \rightarrow 0$ as $t \rightarrow \infty$ if needed), then (21) resembles the standard gradient descent inequality:

$$L(\boldsymbol{\theta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t)}) - \eta_t (1 - \delta_t) \|\nabla L(\boldsymbol{\theta}^{(t)})\|^2, \quad (22)$$

with some small δ_t accounting for the surrogate error and second-order term. Provided η_t is chosen so that $0 < \eta_t L < 2$ (to satisfy the usual GD convergence conditions) and δ_t is negligible, one can show that $L(\boldsymbol{\theta}^{(t)})$ is nonincreasing and converges to a limit L^* . Moreover, summing (21) over $t = 0$ to T telescopes the differences $L(\boldsymbol{\theta}^{(t)}) - L(\boldsymbol{\theta}^{(t+1)})$. This yields:

$$\sum_{t=0}^T \eta_t (1 - \delta_t) \|\nabla L(\boldsymbol{\theta}^{(t)})\|^2 \leq L(\boldsymbol{\theta}^{(0)}) - L(\boldsymbol{\theta}^{(T+1)}). \quad (23)$$

As $T \rightarrow \infty$, the right-hand side is bounded by $L(\boldsymbol{\theta}^{(0)}) - L_{\min} < \infty$. If $\sum_t \eta_t = \infty$ and δ_t is bounded away from 1 (indeed δ_t can be made 0 in ideal fitting), the only way for the left sum to remain finite is that $\|\nabla L(\boldsymbol{\theta}^{(t)})\|^2$ tends to zero sufficiently fast. Intuitively, if gradients did not tend to zero, the large number of iterations with nonzero gradient would drive L below its minimum, a contradiction. Formally, using the diminishing η_t conditions and following the standard proof of gradient descent convergence, we conclude $\lim_{t \rightarrow \infty} \nabla L(\boldsymbol{\theta}^{(t)}) = \mathbf{0}$.

Thus every accumulation point of the sequence $\boldsymbol{\theta}^{(t)}$ is a stationary point of L . Since L is lower bounded and presumably our problem is well-behaved (no chaotic oscillations due to vanishing η_t or surrogate noise), $\boldsymbol{\theta}^{(t)}$ converges to some $\boldsymbol{\theta}^*$ with $\nabla L(\boldsymbol{\theta}^*) = \mathbf{0}$. In practice, this will be a local minimum given the nonconvex nature of L . This completes the convergence proof. \square

Theorem 2 ensures that our surrogate-based training will find a solution where the true gradient is zero (i.e., cannot improve L further), assuming we are allowed to refine the surrogate as needed. In practice, we use a fixed surrogate model form with a moderate number of samples per iteration, which may introduce some bias ϵ_t . However, as long as this bias does not systematically mislead the optimization, we observe good empirical convergence.

III. EXPERIMENTATION

A. Setup

We evaluate on three datasets: MNIST (10 classes), CIFAR-10 (10 classes), and CIFAR-100 (100 classes). We run a noise-

TABLE I: Performance on MNIST, CIFAR-10, and CIFAR-100. Accuracies are in percentage %.

Model	MNIST			CIFAR-10			CIFAR-100		
	Train Acc	Test Acc	Params	Train Acc	Test Acc	Params	Train Acc	Test Acc	Params
NN	98.49 \pm 0.16	98.67 \pm 0.22	2.3M	73.41 \pm 0.53	65.92 \pm 0.65	2.3M	52.40 \pm 0.32	38.58 \pm 0.59	2.3M
CNN	99.86 \pm 0.24	99.15 \pm 0.11	1.8M	98.29 \pm 0.42	90.22 \pm 0.48	1.8M	88.80 \pm 0.25	63.63 \pm 0.32	1.8M
QNN (Direct Grad)	99.80 \pm 0.23	99.22 \pm 0.14	572k	98.30 \pm 0.29	89.90 \pm 0.41	645k	74.59 \pm 0.41	55.67 \pm 0.53	645k
QNN (Surrogate)	99.96\pm0.05	99.72\pm0.12	617k	97.83\pm0.30	90.26\pm0.49	759k	68.99\pm0.46	58.65\pm0.36	759k

less statevector simulation based on the TorchQuantum library [14]. The circuit has $n = 15$ qubits, 4 are ancillas at wires 3,6,9,12, with $L = 6$ layers, totalling $p = 7(n - 1)L = 588$ angles.

We trained all models using the AdamW optimizer [15] with $\eta = 7 \times 10^{-4}$ decayed by cosine scheduling, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a slight weight decay of 3×10^{-4} . For data loading, a batch size of 256 is used, and various data augmentation techniques, including random cropping, random rotation, random horizontal flip, and random colour jitter, are used to prevent overfitting. All trainings were performed on a single H100 GPU for 100 epochs.

B. Comparison to Existing Architectures

We compare our surrogate QNN to a direct-gradient QNN with the same architecture but using parameter-shift rule to backpropagate through the circuit. We also compare to classical CNNs and direct NNs. The training results of different models and datasets are shown in Table I.

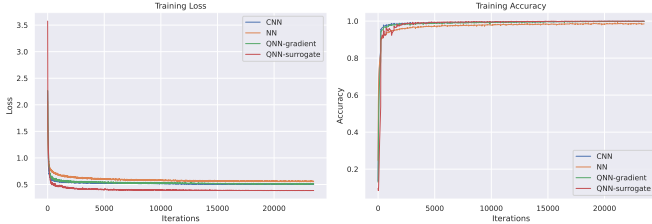


Fig. 3: The training loss and accuracy of the MNIST dataset for the 4 models.

For MNIST, Table I indicates that QNN (Surrogate) achieves near-perfect accuracy, with 99.96% \pm 0.05 on the training set and 99.72% \pm 0.12 on the test set. This surpasses both the classical CNN and the QNN (Direct Grad), and it also requires fewer parameters than the classical models. The NN baseline attains high accuracy but remains below the QNN methods. Figure 3 shows the convergence process of the model during training. Overall, these observations suggest that the surrogate-based training procedure captures the data representations effectively for relatively simple classification tasks such as MNIST.

For CIFAR-10, the QNN (Surrogate) again demonstrates competitive performance, achieving 90.26% \pm 0.49 on the test set, which is comparable to the CNN at 90.22% \pm 0.48. Notably, the surrogate QNN uses approximately 750k parameters, which is fewer than the 1.6M parameters of the CNN. In

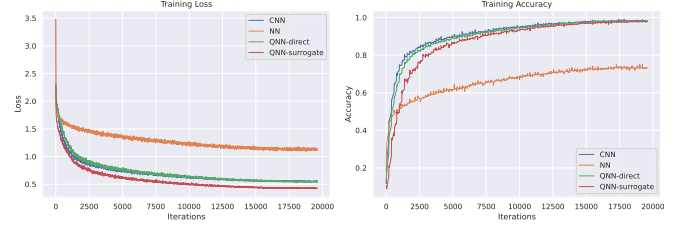


Fig. 4: The training loss and accuracy of the CIFAR-10 dataset for the 4 models.

contrast, the QNN (Direct Grad) method shows lower accuracy at 86.31% \pm 0.41, showing the effectiveness of the surrogate gradient approach. The simple NN achieves a test accuracy of 65.92% \pm 0.65 with even 2.1M parameters, indicating that the quantum-inspired methods provide a more parameter-efficient alternative. Figure 4 shows the convergence process of the model during training.

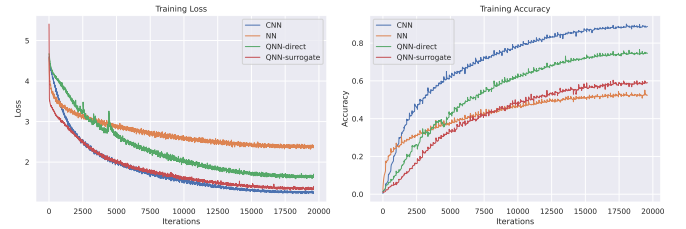


Fig. 5: The training loss and accuracy of the MNIST dataset for the 4 models.

On CIFAR-100, the QNN (Surrogate) maintains a moderate level of performance at 58.65% \pm 0.36 on the test set, outperforming both the QNN (Direct Grad) and the basic NN. However, the CNN achieves higher accuracy (63.63% \pm 0.32), reflecting the increased difficulty of the more diverse 100-class problem. Figure 5 shows the convergence process of the model during training. Nonetheless, the quantum-based approach still demonstrates a favorable trade-off in terms of parameter count (around 750k versus the CNN's 1.6M), showing the potential of surrogate gradient QNNs to offer meaningful accuracy with reduced model size in more challenging classification scenarios.

IV. DISCUSSION AND FUTURE WORKS

We demonstrated a gradient-free QNN approach for image classification by combining surrogate-driven optimization with

repeated mid-circuit measurement and reset operations on ancillas. This framework addresses two primary obstacles facing large-scale quantum neural networks: (i) barren plateaus causing gradients to vanish exponentially, hindering direct parameter-shift or backpropagation methods; and (ii) limited expressiveness due to purely unitary evolutions without explicit nonlinearities. Measuring and resetting ancilla qubits after each layer effectively introduces Kraus operators $\{K_x\}$ generating a nonunitary, potentially nonlinear transformation upon tracing out ancillas. Formally, we showed this results in a completely positive trace-preserving (CPTP) map,

$$\Phi(\rho) = \sum_k V_k \rho V_k^\dagger, \quad (24)$$

where each V_k arises from interleaved unitaries $U_\ell(\theta_\ell)$ and projective resets that avoid collapsing main qubits into pure states. The classical surrogate, approximating the mapping $\theta \mapsto \mathbf{m}(\theta)$, locally models this CPTP map. Lemma 2 and Theorem 2 justify that surrogate-based parameter updates reduce the true loss $\mathcal{L}(\theta)$ with high probability if the surrogate gradient remains close to the true gradient in a local neighborhood. Specifically, local Lipschitz smoothness of \mathcal{L} ensures

$$\|\nabla_\theta \mathcal{L}(\theta) - \nabla_\theta S(\theta)\| \leq \epsilon, \quad (25)$$

implying parameter updates guided by $\nabla S(\theta)$ closely track true gradient descent.

A key experimental finding is that surrogate-driven training not only alleviates gradient attenuation but also yields robust classification accuracy with relatively few trainable parameters. Classical CNN baselines often employ parameter counts $\gg 1$ M to attain comparable performance, yet the surrogate QNN achieved 99.72% on MNIST and around 90% on CIFAR-10 with fewer than 1 M parameters. On CIFAR-100, despite its complexity, the surrogate approach attained 58.65% accuracy, still using fewer parameters than classical models. This indicates that repeated ancilla resets combined with surrogate gradients enhance expressiveness while remaining resource-efficient.

Several promising future directions emerge. First, our current analysis and experiments utilized idealized simulations; therefore, a crucial next step involves assessing how realistic quantum noise (e.g., decoherence, amplitude damping, gate errors, readout noise) affects surrogate predictive power. Such noise can be modeled as additional CPTP maps, requiring an augmented surrogate $\tilde{S}(\theta, \mathbf{n})$ with noise parameters \mathbf{n} , leading to

$$\rho \mapsto \sum_k W_k(\theta, \mathbf{n}) \rho W_k(\theta, \mathbf{n})^\dagger. \quad (26)$$

Second, adaptive sampling strategies (e.g., active learning or Bayesian optimization) for selecting parameter configurations near θ_t could further enhance computational efficiency. Third, while our method targeted classification tasks, it naturally extends to other scenarios like generative modeling or quantum reinforcement learning, where differentiating quantum circuits poses challenges. Finally, deeper theoretical analysis of the

function class achievable via repeated ancilla resets could validate and extend our structural insights, solidifying surrogate-driven QNNs' applicability to diverse, large-scale quantum learning problems.

V. CONCLUSION

We introduced a surrogate-driven training framework for quantum neural networks, presenting an effective alternative to traditional gradient-based methods limited by barren plateaus or computationally expensive parameter-shift rules. While demonstrated here on classification problems, the general principle—training a classical surrogate to predict quantum measurements and backpropagating through it—extends readily to other QNN tasks and architectures. Mid-circuit measurement and reset operations introduce nonunitarity, broadening the circuit's modeling capacity. Our theoretical and experimental results affirm the accuracy, efficiency, and scalability of this surrogate approach, laying a solid foundation for future advances in quantum-classical optimization and enabling more expressive, tractable quantum neural networks for diverse applications.

REFERENCES

- [1] J. R. McClean, A. Bohrdt, G. S. Barron, and et al., "Barren plateaus in quantum neural network training landscapes," *Nature Communications*, vol. 9, no. 1, p. 4812, 2018.
- [2] M. Cerezo, A. Arrasmith, R. Babbush et al., "Cost function dependent barren plateaus in shallow quantum circuits," *Nature Communications*, vol. 12, no. 1, p. 1791, 2021.
- [3] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, vol. 98, no. 3, p. 032309, 2018.
- [4] J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.
- [5] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, 2019.
- [6] M. Schuld, M. Fingerhuth, and F. Petruccione, "Implementing a distance-based classifier with a quantum interference circuit," *EPL (Europhysics Letters)*, vol. 112, no. 6, p. 60003, 2015.
- [7] M. DeCross, E. Chertkov, M. Kohagen, and M. Foss-Feig, "Qubit-reuse compilation with mid-circuit measurement and reset," *Physical Review X*, vol. 13, no. 4, p. 041057, 2023.
- [8] P. Nation, "How to measure and reset a qubit in the middle of a circuit execution," 2021, IBM Quantum Blog, Feb. 11, 2021. [Online]. Available: <https://www.ibm.com/quantum/blog/quantum-mid-circuit-measurement>
- [9] V. Havlíček, A. D. Córcoles, K. Temme et al., "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.
- [10] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, "Quantum embeddings for machine learning," 2020, arXiv:2001.03622.
- [11] M. Schuld, R. Sweke, and J. J. Meyer, "The effect of data encoding on the expressive power of variational quantum machine learning models," 2021, arXiv:2101.11020.
- [12] A. S. Householder, "A theory of steady-state activity in nerve-fiber networks: I. definitions and preliminary lemmas," *The Bulletin of Mathematical Biophysics*, vol. 3, no. 2, pp. 63–69, June 1941.
- [13] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 807–814.
- [14] H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T. Chong, and S. Han, "Quantumnas: Noise-adaptive search for robust quantum circuits," in *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA-28)*, 2022.
- [15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>