A PID Controller Approach for Adaptive Probability-dependent Gradient Decay in Model Calibration

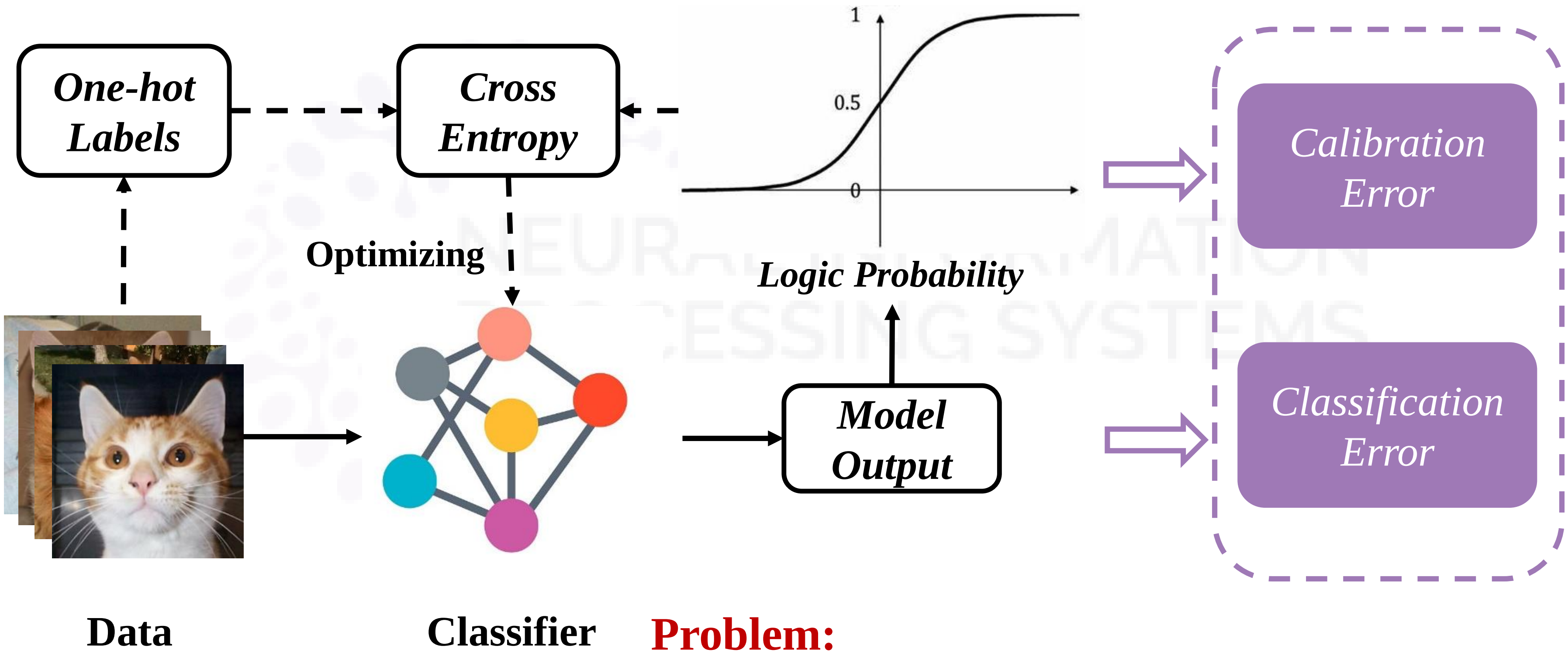

Siyuan Zhang **Linbo Xie**

School of Internet of Things Engineering

Jiangnan University, Wuxi, China

The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

Background



Problem:

How can <u>*Calibration Error*</u> and <u>*Classification Error*</u> be consistently optimized?

Motivation

Parametric Softmax: $J = -\log \frac{e^{z_c/\tau}}{\sum_{i \neq c} e^{z_i/\tau} + \beta e^{z_c/\tau}}$

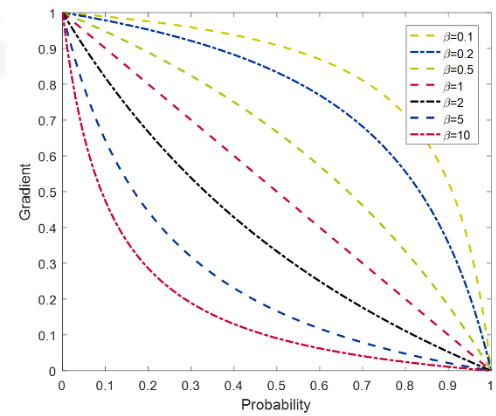
Temperature

It can be regard as a soft margin in output space.

Proposed probability-dependent gradient decay rate

Introduce probabilistic output $p_i = \frac{e^{z_i}}{e^{z_1+\cdots+e^{z_m}}}$ as an intermediate variable. Then we obtain:

$$\frac{\partial J}{\partial z_i} = \begin{cases} -\frac{1-p_c}{1+(\beta-1)p_c}, i = c \\ \frac{p_i}{1+(\beta-1)p_c}, i \neq c \end{cases}$$

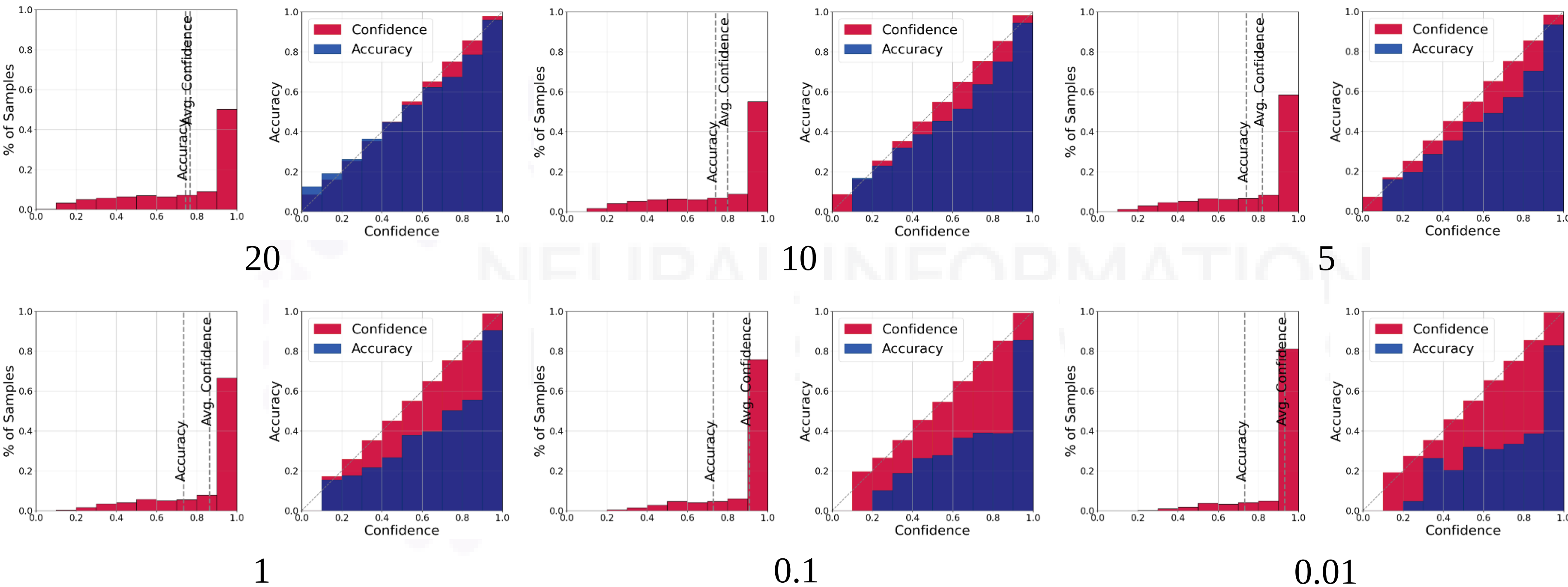


Finding:

Probability-dependent gradient decay rate is closely correlated with *model calibration*.

Motivation

Confidence and reliability diagrams with different gradient decay rate



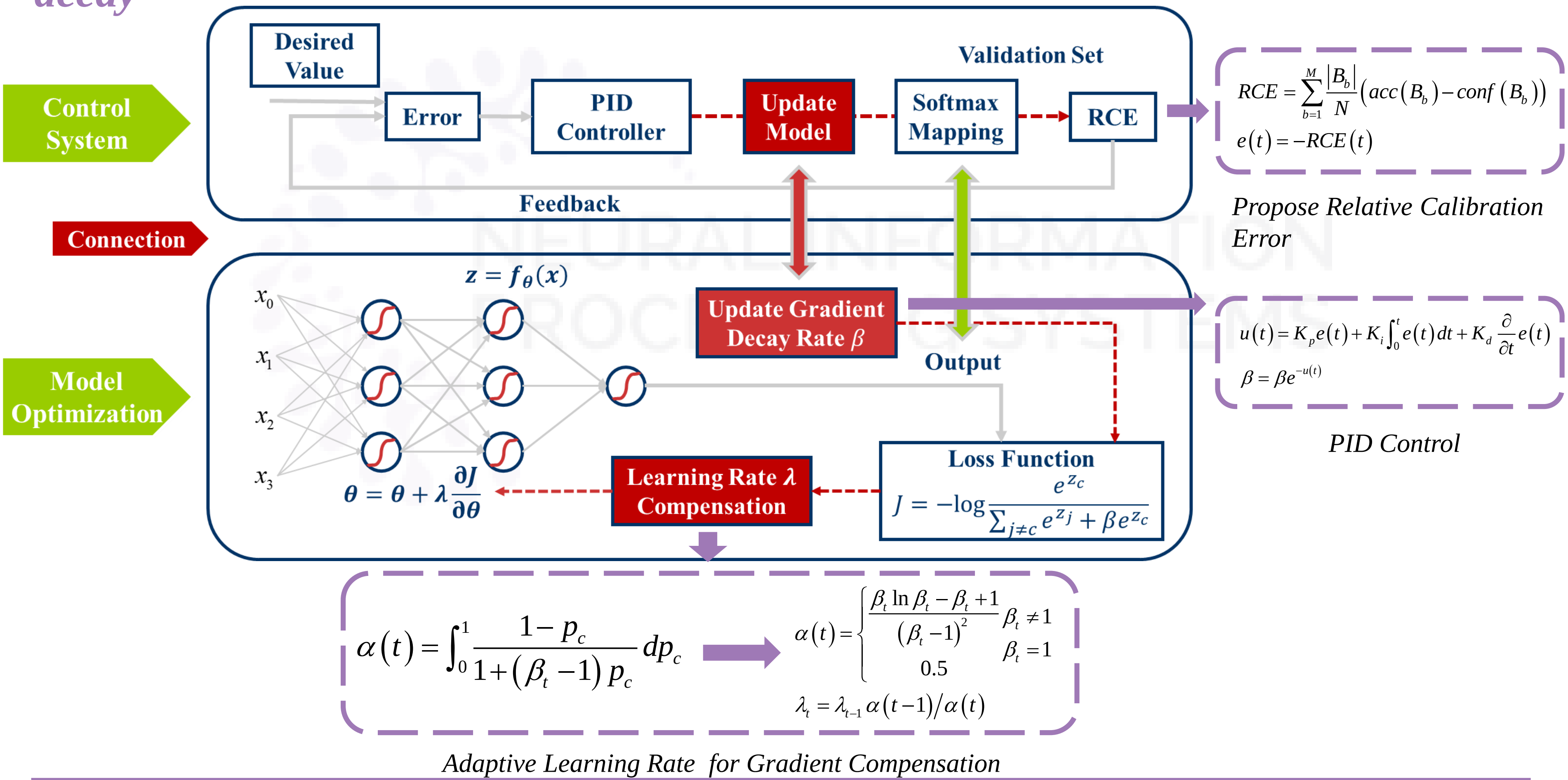
Finding:

A small decay rate results in a curriculum learning sequence.

Our results show *a negative correlation* between *the gradient decay rate* with increasing instance-level probability and the *overall confidence distribution*.

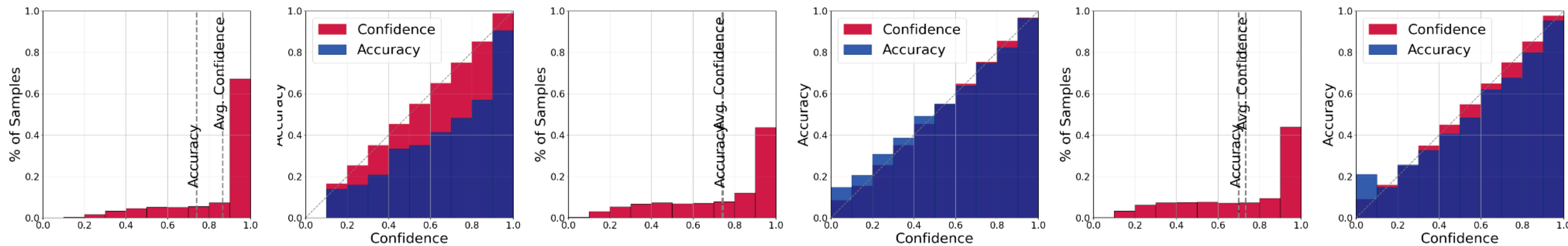
Methodology

The framework of PID controller-based adaptive probability-dependent gradient decay

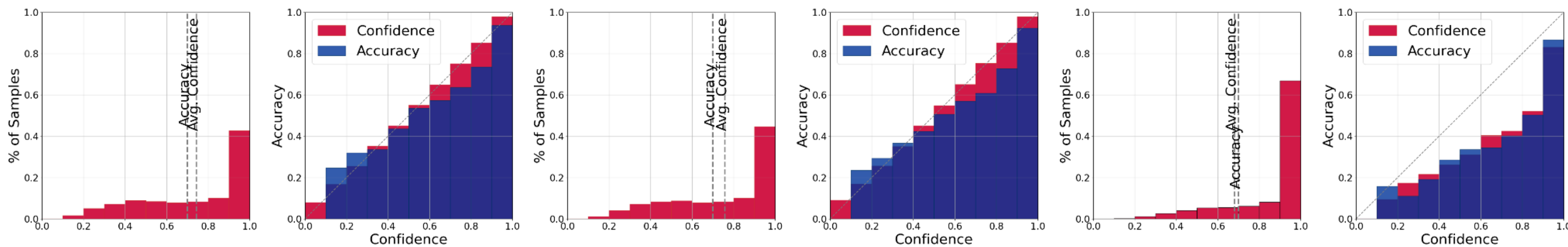


Some Results

1. Calibration performance with other post-processing calibration methods



Uncalibrated Ours Temperature Scaling



Matrix Scaling Vector Scaling Hist. Bin.

Conclusion:

The experimental findings underscore the effectiveness of our approach by dynamically adjusting the gradient decay rate during the model optimization.

Some Results

2. Performance of consistent optimization in supervised learning

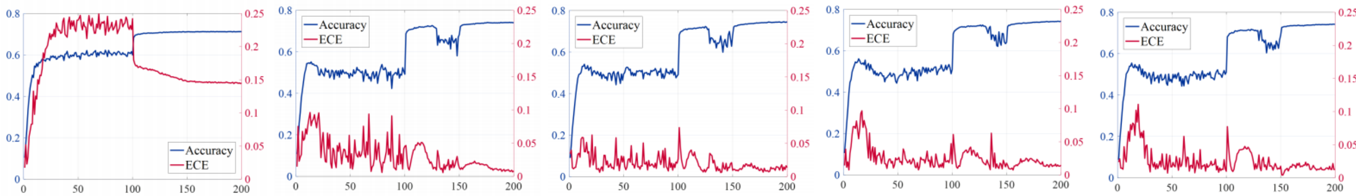
Methods	Models	CIFAR-10				CIFAR-100			
		ACC (%)	ECE	MCE	AdaECE	ACC (%)	ECE	MCE	AdaECE
Softmax	ResNet18	93.7±0.39	0.041±0.010	0.281±0.076	0.042±0.013	73.6±0.29	0.160±0.026	0.344±0.048	0.160±0.026
	ResNet35	93.9±0.39	0.054±0.015	0.300±0.083	0.054±0.016	73.8±0.30	0.172±0.022	0.351±0.077	0.172±0.023
	VGG16	92.1±0.41	0.066±0.022	0.339±0.091	0.068±0.023	69.2±0.26	0.233±0.054	0.476±0.112	0.236±0.053
Cosface	ResNet18	93.9±0.45	0.053±0.011	0.352±0.072	0.055±0.013	74.2±0.51	0.185±0.046	0.501±0.162	0.183±0.050
	ResNet35	**95.6**±0.42	0.048±0.012	0.317±0.095	0.049±0.011	74.6±0.38	0.181±0.065	0.488±0.127	0.178±0.063
	VGG16	92.7±0.58	0.067±0.019	0.390±0.101	0.068±0.020	71.4±0.52	0.238±0.081	0.567±0.125	0.233±0.085
Center loss	ResNet18	94.5±0.41	0.038±0.009	0.337±0.075	0.040±0.008	74.1±0.30	0.082±0.013	0.222±0.071	0.085±0.015
	ResNet35	95.5±0.51	0.043±0.010	0.280±0.099	0.045±0.012	74.2±0.31	0.098±0.031	0.250±0.096	0.101±0.030
	VGG16	**93.1**±0.41	0.034±0.009	0.349±0.083	0.034±0.010	**72.1**±0.37	0.216±0.042	0.472±0.104	0.231±0.045
DCA	ResNet18	91.9±0.32	0.020±0.006	0.156±0.038	0.022±0.007	72.1±0.25	0.047±0.011	0.156±0.024	0.049±0.012
	ResNet35	92.3±0.43	0.035±0.012	0.186±0.046	0.034±0.010	73.1±0.28	0.067±0.021	0.184±0.051	0.066±0.023
	VGG16	90.7±0.28	0.027±0.008	0.255±0.078	0.027±0.008	70.9±0.37	0.133±0.028	0.269±0.059	0.141±0.032
Ours	ResNet18	**95.0**±0.41	**0.007**±0.002	**0.078**±0.021	**0.008**±0.001	**74.3**±0.43	**0.006**±0.002	**0.068**±0.018	**0.007**±0.002
	ResNet35	95.6±0.51	**0.009**±0.002	**0.089**±0.012	**0.010**±0.003	**75.4**±0.39	**0.011**±0.003	**0.063**±0.011	**0.014**±0.002
	VGG16	92.6±0.35	**0.011**±0.002	**0.083**±0.031	**0.012**±0.004	71.9±0.35	**0.028**±0.008	**0.044**±0.003	**0.030**±0.010

Conclusion:

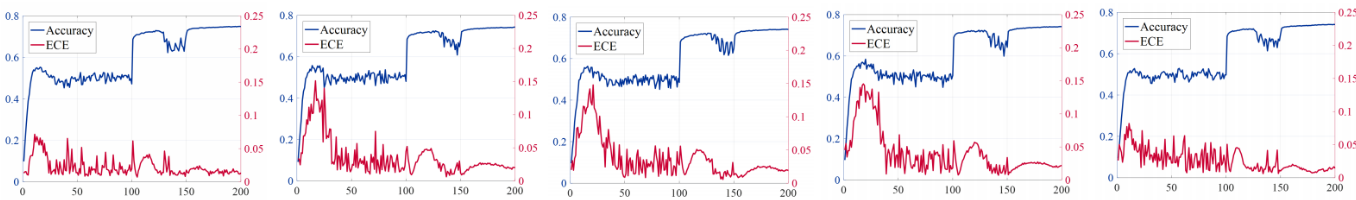
Our proposed PID-based method with variable gradient decay rate ensures both model accuracy and calibration.

Some Results

3. Ablation experiments and analysis for PID controller



(a) P:0 I:0 D:0 (b) P:1 I:1 D:1 (c) P:1 I:2 D:1 (d) P:1 I:0.1 D:1 (e) P:1 I:0.01 D:1



(f) P:1 I:0.01 D:2 (g) P:1 I:0.5 D:0.1 (h) P:1 I:0.5 D:0.01 (i) P:0.5 I:0.5 D:0.1 (j) P:5 I:0.5 D:0.1

Conclusion: Model calibration is robust to the choice of PID parameters.

4. Ablation experiments of different optimizer

SGD	Adam	PID Controller Approach	Gradient Compensation	Accuracy	ECE	AdaECE
✓	-	-	-	73.8%	0.172	0.172
✓	-	✓	-	72.5%	0.022	0.023
-	✓	✓	-	63.5%	0.023	0.024
✓	-	✓	✓	74.7%	0.012	0.013

Conclusion: Adaptive learning rate for gradient compensation can significantly improve the performance of PID control-based calibration.

Thank you for your attention! ☺

For more details, please refer to our paper !