NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026

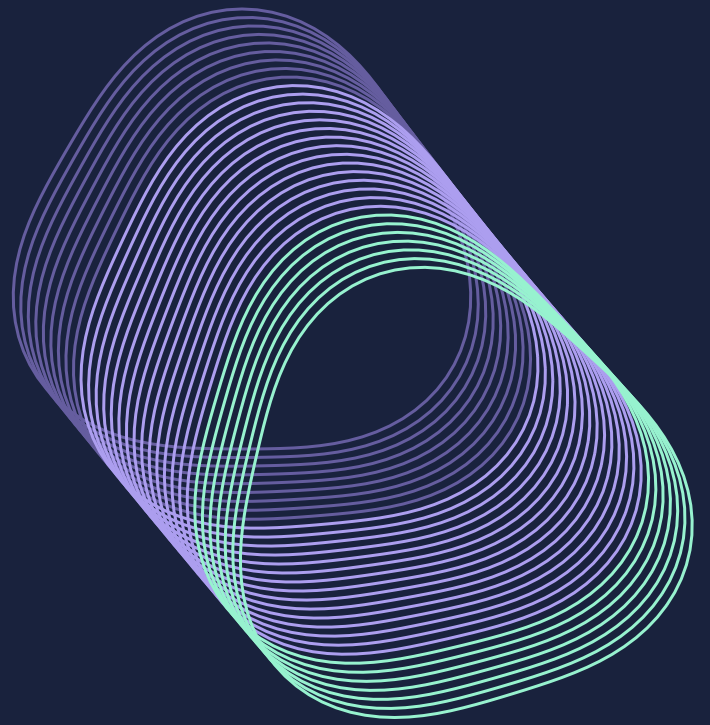
Beyond FLOPs:

Train-Full, Deploy-Partial Multi-Exit Inference via Selective Lightweight IC Ensemble

Bitchan Eom[1], Eunchan Kim[1,2]*

[1] Department of Artificial Intelligence, Hanyang University
[2] Department of Information Systems, Hanyang University
Email: {ubc573, eckim}@hanyang.ac.kr

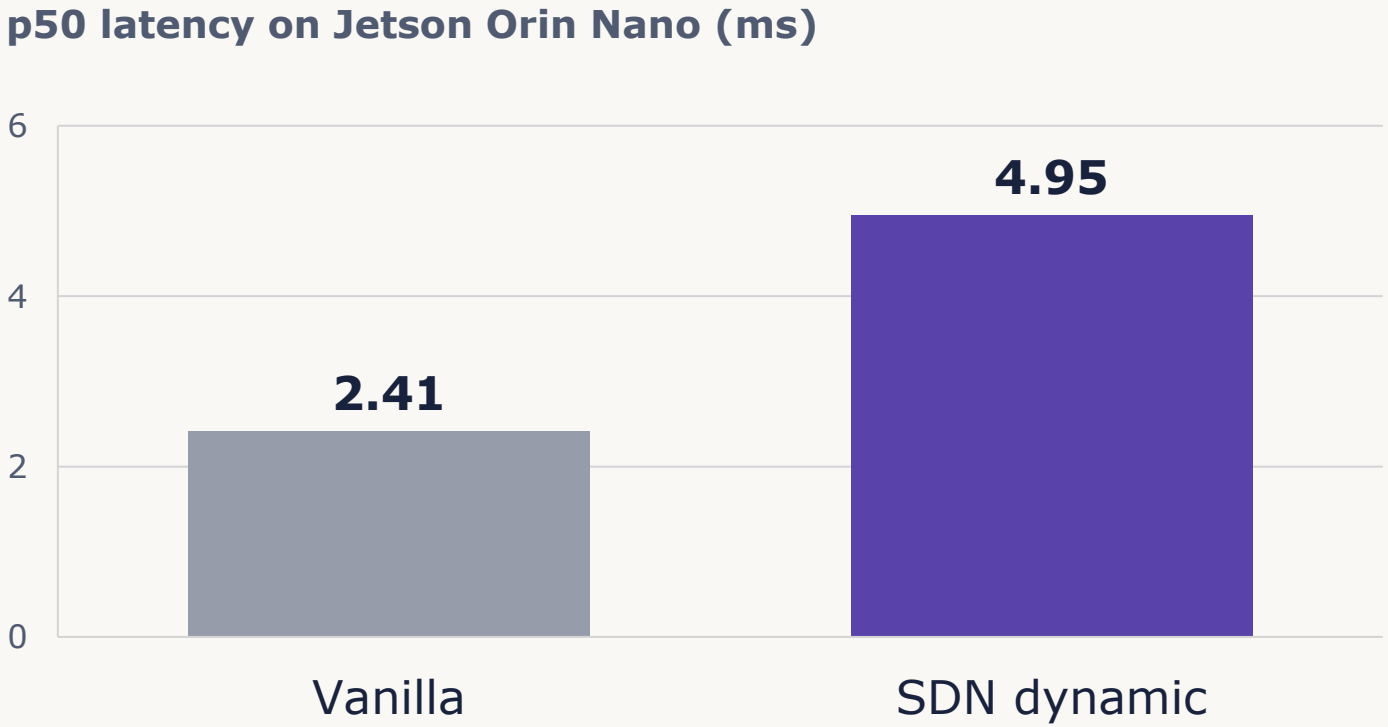
*Corresponding author



SLIENet

MULTI-EXIT INFERENCE

Early exits doubled median latency



2.1×

the latency of vanilla ResNet-56

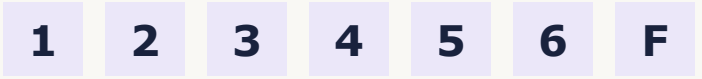
SDN can skip layers, but our TensorRT implementation launches 1–7 engines per image. Synchronization and copies offset the FLOP savings.

Table 4. TensorRT FP16, fixed seed-0 checkpoint, locked clocks, 10 runs of 10,000 CIFAR-100 images.

One checkpoint serves several depth budgets

01 **Train the full network**

Train six internal classifiers (ICs) and a final head.



02 **Select the IC subset**

At depth 5, calibration selects IC4 and IC5.



03 **Deploy one static engine**

Compile the retained backbone and selected heads.



SLIE4, SLIE5 and SLIE6 share weights. Each budget compiles to its own TensorRT FP16 engine.

Why average only IC4 and IC5?

During training

The final head teaches each IC.
A stop-gradient blocks the distillation gradient at the teacher output.

At deployment

SLIE5 stops at depth 5. It averages the class probabilities from IC4 and IC5 with equal weights.

CIFAR-100 top-1 accuracy	
All ICs through IC5	**67.38%**
Single IC5	**69.49%**
Selected IC4 + IC5	**69.81%**

Table 3. Mean accuracy across three seeds. A held-out calibration split determines which heads to average.

SLIE5 improves accuracy while using less energy

+1.2 pp

Top-1 accuracy

68.66% to 69.81%

−30%

p50 latency

2.41 to 1.69 ms

−40%

Energy per image

13.02 to 7.81 mJ

Same locked-clock session

Vanilla ResNet-56

2.41 ms

SDN dynamic

4.95 ms

SLIE5

1.69 ms

Vs. vanilla. Accuracy averages three training seeds (Table 3). Device metrics average 10 locked-clock Jetson runs (Table 4).

A dedicated student has a small accuracy edge

Model	Top-1 accuracy	MACs (M)
Vanilla ResNet-56	68.66%	125.8
SLIENet, single IC5	69.49%	97.5
SLIE5	**69.81%**	**97.6**
Distilled truncated	70.15%	97.5

+0.34 pp

for the distilled student at almost identical compute.

Training that student fixes its depth. One SLIENet checkpoint yields SLIE4, SLIE5 and SLIE6 at 79.7M, 97.6M and 116.6M MACs.

Table 3. ResNet-56 on CIFAR-100. Accuracy averages three seeds for vanilla and SLIENet, and five for the distilled student.

ZTW gains 0.44 points at higher device cost

Deployed model	Top-1	p50 latency	Energy / image	Engine calls
SLIE6	**70.91%**	**1.90 ms**	**11.56 mJ**	**1**
ZTW	71.35%	4.90 ms	32.57 mJ	6

2.6× p50 latency

2.8× energy per image

Scope of the comparison

We measured CNNs on Jetson Orin Nano with TensorRT FP16. Full ImageNet-scale models, Transformers and a custom fused ZTW engine remain untested.

Table 4. Accuracy uses the deployed seed-0 checkpoint. Latency and energy average 10 runs in the same locked-clock session.

What determines the cost of an early exit?

01 **Fewer FLOPs can still mean higher latency.**

SDN dynamic took 4.95 ms versus 2.41 ms for vanilla.

02 **One training run supports several budgets.**

Each selected depth compiles to one static engine.

03 **Selecting the heads improves the trade-off.**

SLIE5 beat vanilla on accuracy, latency and energy in our test.

Paper · OpenReview

Code · GitHub

Bitchan Eom and Eunchan Kim / Hanyang University