Improving GreedyViG for CIFAR Image Classification

Agam Harpreet Singh (B23CM1004) — Ishan Shah (B23CM1050)

YouTube: https://www.youtube.com/watch?v=HXRYf8B09wQ

GitHub: https://github.com/Agam77055/GreedyViG-CIFAR100

Abstract

We present an improved adaptation of GreedyViG—a hybrid CNN-GNN vision architecture using Dynamic Axial Graph Construction (DAGC)—for image classification on the **CIFAR-100** dataset. The baseline model achieved only 51.7% accuracy on CIFAR-100, far below its reported ImageNet performance. Through careful analysis we identified four root causes: a silent graph construction no-op that rendered all graph convolutions inert on 32×32 inputs, a non-differentiable hard connectivity mask, absent channel attention, and a kernel mis-wiring that locked receptive fields at 3×3. We introduce four complementary fixes: a CIFAR-aware stem with corrected graph loop bounds, a learnable threshold parameter α with a soft sigmoid mask, Squeeze-and-Excitation (SE) channel attention with a corrected 5×5 depthwise kernel, and dual max+mean graph aggregation. The improved model reaches **72.90% top-1 accuracy** in 100 epochs, a **+21.2 pp absolute gain**.

1. Base Paper Overview

1.1 Problem & Motivation

Vision Graph Neural Networks (ViGs) represent image patches as graph nodes, enabling richer non-local interactions than CNNs. However, canonical K-Nearest Neighbour (KNN) graph construction is quadratic in the number of patches—a severe efficiency bottleneck. Static graph methods (e.g., MobileViG's SVGA) avoid KNN at the cost of input-independent graphs, sacrificing the core benefit of dynamic connectivity.

1.2 Proposed Solution: DAGC & GreedyViG

Dynamic Axial Graph Construction (DAGC) estimates the global distance distribution—mean μ and standard deviation σ—from a lightweight quadrant-flip comparison, then rolls axially to connect only nodes whose distance satisfies $d < \mu - \sigma$. Key properties:

- **Sub-quadratic cost** — no all-pairs KNN; comparisons scale linearly.
- **Input-adaptive** — the edge mask changes per image, unlike SVGA's fixed graph.
- **Reshaping-free** — operates on 4-D feature maps directly.

GreedyViG pairs DAGC blocks with MBConv blocks across four stages, combining local (MBConv) and global (DAGC) processing at every resolution. Conditional positional encoding (CPE) and max-relative graph convolution round out the design.

Table 1: GreedyViG vs. comparable ViG models on ImageNet-1K.

Model	Params	GMACs	Top-1
GreedyViG-S	12.0M	1.6	81.1%
GreedyViG-M	21.9M	3.2	82.9%
GreedyViG-B	30.9M	5.2	83.9%
PViG-B	92.6M	16.8	83.7%
PViHGNN-B	94.4M	18.1	83.9%

1.3 Key Results (ImageNet-1K)

GreedyViG-B matches PViHGNN-B accuracy with **67% fewer parameters** and **71% fewer GMACs**, validating DAGC's efficiency across classification, detection, and segmentation benchmarks.

2. Our Approach & Contributions

2.1 Identified Issues in the Baseline

Running the unmodified model on CIFAR-100 revealed four compounding failures:

- **Graph loop no-op.** The axial loop runs `range(0, H, K)`. After the stride-2 stem, every stage satisfies $H = K$ (e.g. $H = K = 8$ in stage 0), so the loop executes only once at $i = 0$. A roll by zero is the identity; the neighbour difference and the entire graph output are identically zero. The model reduced to a pure-CNN baseline.
- **Non-differentiable hard mask.** The binary threshold $\mathbf{1}[d < \mu - \sigma]$ has zero gradient everywhere; connectivity cannot be refined by backpropagation.
- **No channel attention.** DepthWiseSeparable blocks treated all channels equally, with no mechanism to suppress uninformative feature maps.
- **Silent kernel mis-wiring.** The configurable `kernel` argument was never forwarded to the depthwise conv; padding was hard-coded to 1, locking the receptive field at 3×3 regardless of configuration.

2.2 Fix 1 — CIFAR Adaptation & Graph Loop Correction

Small-input stem. The original stride-2×2 stem collapses 32^2 inputs to 8^2. We replace both strides with stride-1, preserving full 32×32 resolution into the backbone. Hop distances are updated to $K = [4, 2, 1, 1]$ across stages; feature maps progress $32 \to 16 \to 8 \to 4$. CIFAR-100-specific normalisation statistics

replace the hardcoded ImageNet constants.

Range loop fix. The axial loop is corrected to `range(K, H, K)`, starting at the first non-trivial roll offset. With the updated feature-map sizes, stage 0 now performs 7 meaningful row-wise and 7 column-wise neighbour comparisons per forward pass instead of zero.

2.3 Fix 2 — Learnable Adaptive Graph Threshold

The original fixed rule $d < \mu - \sigma$ is replaced by a learnable soft threshold:

$$\tau = \mu - \alpha\,\sigma, \qquad m = \sigma\big(\tfrac{\tau - d}{\sigma}\big)$$

where α is a per-block scalar parameter initialised to 1 (recovering the original rule at initialisation) and $\sigma(\cdot)$ is the sigmoid function. This makes two improvements: (i) α is trained by backpropagation, allowing each block to learn its optimal connectivity sparsity; (ii) the soft sigmoid mask provides non-zero gradients near the threshold boundary, enabling end-to-end differentiable graph construction.

2.4 Fix 3 — SE Channel Attention & Kernel Correction

A Squeeze-and-Excitation block is inserted after the depthwise conv in every local DepthWiseSeparable block. Global average pooling compresses the $4C$-channel expanded features to a C-dim descriptor; a FC→ReLU→FC→Sigmoid pathway learns per-channel importance weights. The kernel mis-wiring is corrected simultaneously: padding is set to `kernel // 2` and `kernels=5` is forwarded to the depthwise conv, widening the local receptive field on 32×32 feature maps.

2.5 Fix 4 — Dual Aggregation (Max + Mean)

Inside the DAGC rolling loop we now accumulate both max and mean statistics over masked neighbour differences:

$$\mathbf{x}_j = \mathbf{x}_{\max} + \mathbf{x}_{\text{mean}}$$

The combined descriptor pairs peak-signal sensitivity (max) with contextual consistency (mean). Because both are computed in the same loop pass, this fix adds **zero extra parameters** and negligible compute overhead.

2.6 Training Strategy

- Mixup ($\alpha = 0.4$) and CutMix ($\alpha = 0.5$) for label-smoothed interpolation.
- RandAugment (m=7, std=0.5) for appearance diversity.
- Random Erasing ($p = 0.1$) to regularise local feature reliance.
- Model EMA (decay=0.9992) for stable late-epoch predictions.
- 5-epoch linear warmup + cosine annealing over 100 epochs.

3. Experiments & Results

3.1 Experimental Setup

All experiments use **CIFAR-100** (50K train / 10K test, 32×32 RGB, 100 classes). Both runs train for **100 epochs**, batch size 128, single GPU via `torchrun`. No knowledge distillation is applied.

3.2 Quantitative Results

Table 2: Baseline vs. improved model on CIFAR-100.

Configuration	Top-1 Acc.	Top-5 Acc.
Baseline GreedyViG-S	51.70%	—
GreedyViG-S-CIFAR (ours)	**72.90%**	**93.10%**

Table 3: Estimated per-fix contribution.

Fix	Params	Est. Gain
Stem + loop + K fix	≈ 0	+10–13 pp
Learnable threshold (α)	≈ 0	+2–4 pp
SE + kernel fix (5×5)	$2C^2/r$	+2–3 pp
Dual aggregation	0	+2–3 pp
Training augmentation	0	+1–2 pp

The final model achieves a **+21.21 pp absolute improvement** over the baseline with only 23 additional minutes of training time (3h 09m → 3h 32m).

3.3 Key Insights

- **The graph was entirely silent.** The range loop bug meant GreedyViG operated as a pure MBConv network on CIFAR-100. Fixing it alone accounts for the majority of the accuracy recovery.
- **Differentiable connectivity matters.** Replacing the hard mask with a learnable soft threshold lets each block tune its own sparsity level through gradient descent.
- **Dataset-aware design is essential.** CIFAR-specific normalisation and a stride-1 stem prevent information destruction that would dominate all other improvements.
- **Aggregation diversity is free.** Dual max+mean aggregation requires no new parameters yet enriches the neighbourhood descriptor significantly.

Conclusion. We demonstrated that GreedyViG's DAGC graph construction, while highly effective on ImageNet, required careful re-engineering for CIFAR-100. A critical loop bug silently disabled all graph convolutions; fixing it alongside three complementary improvements lifted accuracy from 51.7% to 72.9% with negligible overhead. Future work includes applying knowledge distillation from a CIFAR-100 pretrained teacher and exploring adaptive aggregation weights to further improve the model.

References

[1] M. Munir, W. Avery, M. Rahman, and R. Marculescu, *GreedyViG: Dynamic Axial Graph Construction for Efficient Vision GNNs*, CVPR, 2024.

[2] J. Hu, L. Shen, and G. Sun, *Squeeze-and-Excitation Networks*, CVPR, 2018.

[3] K. Han et al., *Vision GNN: An Image is Worth Graph of Nodes*, NeurIPS, 2022.