DAGA · NEURIPS 2026 · MAIN TRACK

Dynamic Attention-Guided Adaptation for Self-Supervised Vision Transformers

Tianjian Zhou[1,†] · **Jie Jiang**[1,†] · **Yishan Li**[1,*] · **Yifei Zhang**[2,*]
[1] NUDT • [2] NWPU • [†] equal contribution • [*] corresponding

IN ONE LINE

Stop learning a new transform for every image. Read the attention this image already produced — the frozen stack computes it anyway.

IMAGENET-1K TOP-1
85.9%
85.90 ± 0.06 · n = 3
Layer-FT parity at 18×

+4.4 SIGNAL
ρ=0.93 SCOPE
1.4% TRAINABLE
0.55× TWO-PASS

NEURAL INFORMATION PROCESSING SYSTEMS

01 Attention already knows what matters

Self-distilled ViTs localise objects in their own attention — static PEFT never reads it

DAGA reads the attention the frozen backbone already computes.

- SSL attention is per-image and object-centric — the frozen stack computes it on every forward pass.
- Adapters, LoRA and prompts are content-agnostic: one learned transform per image, so the signal is computed and then thrown away.
- Not attention refinement: CLIPSelf-style front ends improve the map; DAGA consumes whatever map it receives.
- Reading it costs one cached no-grad pre-forward — computed once per image, reused by every gated block.
- One rule per task family, not per dataset — a fixed four-block placement; the sweep spans only 1.9 points.

The gain is the signal, not extra capacity — same 1.20 M adapter.

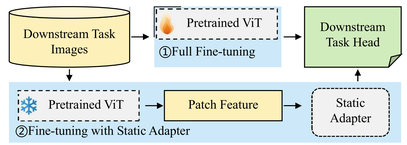


Fig. 1 — Adaptation paradigms (paper Fig. 1).

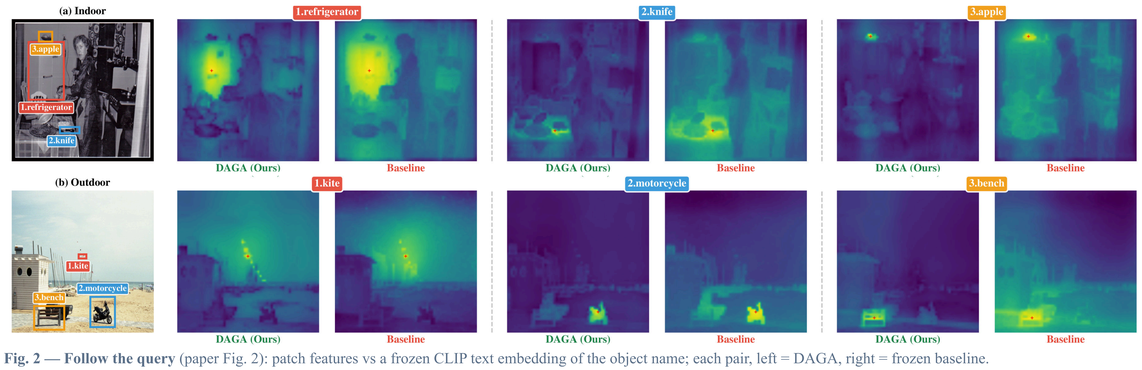


Fig. 2 — Follow the query (paper Fig. 2): patch features vs a frozen CLIP text embedding of the object name; each pair, left = DAGA, right = frozen baseline.

02 Read the attention, encode it once, gate every layer

A shared attention encoder, layer-specific readouts and a gated bottleneck residual — the backbone stays frozen

Gated residual one per inserted block

$$\hat{F}^{(\ell)} = F^{(\ell)} + \alpha \cdot G_c \cdot (1 + \beta\, S) \odot \Psi(F^{(\ell)})$$

four gated layers one shared encoder

What changes a gate per image, per gated block — 1.20 M in total

What does not the frozen backbone and every feature it already computes

- **Ψ** proposes the update: a per-patch bottleneck — one bottleneck per patch, shared by every gated block — **768→107→768**
- **G_c** selects *which* channels keep it — channel gate: which dimensions keep the update — **∈ [0,1]^D**
- **S** selects *where* it lands — spatial gate upsampled to the patch grid — **14 × 14 map**
- **α, β** set how loud the term is — learnable scalars (α ≈ 0.10–0.15, β ≈ 0.1) — **two scalars**
- **zero-init** the branch starts as the exact identity — zero-init up-projection: exact identity at step 0 — **$\hat{F}$ = F at step 0**



WHAT IS LEARNED

Per image, per gated block: one shared encoder, four read-outs — never a per-dataset rule.

1.20 M learned · 86 M frozen · four gated layers.

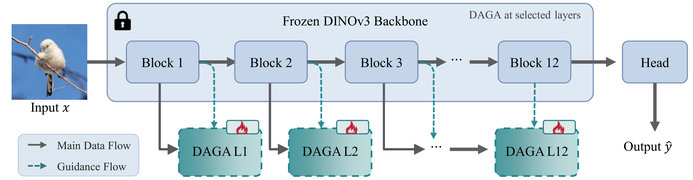


Fig. 3 — DAGA architecture (paper Fig. 4, panels a–b; panel c is the residual above).

A · encode once final-block attention → one 64-channel encoding, cached

B · read out channel gate (128-d) + 14×14 map, per layer

C · gate the residual zero-init bottleneck = exact identity at step 0

Read it as two steps. The frozen backbone's final-block attention is encoded once by a shared CNN; four layer heads turn it into the channel gate G_c and the 14 × 14 map S, and a zero-init bottleneck applies the gated residual.

Backbone DINOv3 ViT-B/16, frozen (86 M) | **Adapter** 1.20 M (1.4 %), r = 107, one gate per gated block | **Placement** blocks {1, 2, 10, 11} | **Protocol** per-method LR sweep, n = 3 seeds

03 What the gates buy

DINOv3 ViT-B/16 frozen (86 M), 224 px · every row re-run by us under one protocol with a per-method LR sweep · n = 3 seeds on IN-1K and dense

ImageNet-1K classification

Top-1, frozen ViT-B/16 · n = 3 seeds except ‡ · bars span 78–86

Method	Train.	Top-1
Linear probe‡	0.1 M	80.90
VPT-Deep	0.6 M	83.05
LoRA	0.3 M	83.45
AdaptFormer	1.2 M	84.38
Full fine-tune‡	85.7 M	84.40
Layer-FT · last‡	7.1 M	84.60
ViT-Adapter	1.5 M	85.07
NOAH‡	0.4 M	85.53
Adapter+‡	1.0 M	85.54
Layer-FT · top 3	21.3 M	85.84
DAGA	**1.20 M**	**85.90**

Parity with Layer-FT in seed noise (±0.06 vs ±0.07) at **18×** fewer parameters; +0.6 over the best-tuned ViT-Adapter (85.3, matched sweep). ‡ single seed.

Dense & fine-grained

Same 1.20 M adapter · default placement, ADE20K matched

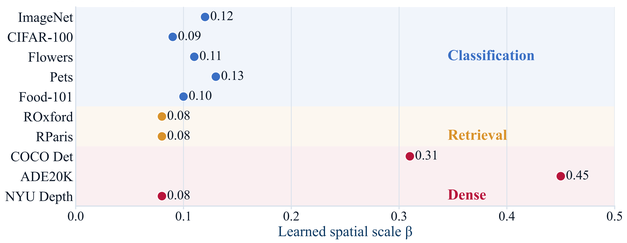
Task	Metric	Result	Δ vs its baseline
ADE20K	mIoU	50.50 ± 0.27	+2.9
COCO	AP	35.62 ± 0.18	+2.1
CUB-200	top-1	89.46	+1.57
FGVC-Aircraft	top-1	85.75	+3.45
NYU Depth	RMSE ↓	0.432	−0.005
ROxford-Med	mAP	86.3	+0.3
ImageNet-C	mean acc	65.3	+0.7

Over the frozen probe the adapter adds **+12.1 mIoU** / **+3.1 AP**; Δ is vs the best static PEFT, except ImageNet-C vs frozen and ROxford-Med vs Layer-FT 86.0. Baselines: COCO vs AdaptFormer 33.5, CUB/Aircraft vs ViT-Adapter, NYU 0.437.

34.1 GFLOPs | **0.55×** throughput | **−39%** train mem | **1.20 M** trainable

What the gate learns

Learned spatial scale β per task family



Task family sets β: 0.08–0.13 global, 0.31–0.45 consumed-map, depth 0.08.

Honest scope

Where the method does not apply

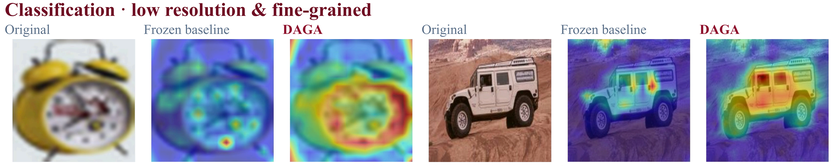
- **Scope.** Self-distilled ViTs only — on MAE, CLIP, DeiT, MoCo v3 the gain is ≤ +0.2 without refinement.
- **Depth.** β = 0.08, RMSE 0.432 vs Layer-FT 0.414 — semantic attention is a poor proxy for geometry.
- **Published dense-PEFT** still wins some cells: LoRand 51.9 mIoU, DeAP 0.425 RMSE at 43 K parameters.
- **Not free.** Two passes (34.1 GFLOPs, 0.55× throughput); the guidance pre-forward is un-mergeable.
- **Seed noise.** n = 3: ±0.06 on IN-1K (±0.07 for Layer-FT) — a within-noise parity claim.

Refine first, then adapt. CLIPSelf on CLIP → **+2.7 / +5.3**; self-distillation on MAE → **+2.3 / +3.6**.

04 Where the gates actually move attention

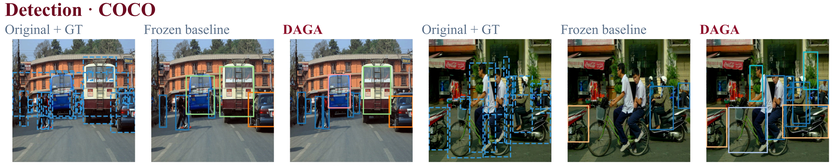
Each trio, left to right: original · frozen baseline · DAGA

Classification · low resolution & fine-grained



Alarm clock · off-road SUV. DAGA concentrates on the object itself, not the background.

Detection · COCO



Urban street · bicycle parking. Sharper, better-localised boxes from the same frozen backbone — +3.1 AP over the frozen probe.

Segmentation · ADE20K



Urban skyline · dining scene. Cleaner boundaries and homogeneous regions — +12.1 mIoU over the frozen probe.

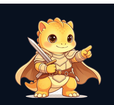
TAKE-HOME MESSAGE

The backbone already knows where to look. DAGA lets its own attention steer the adapter.

1.4% trainable · Layer-FT parity · largest gains where self-distillation makes attention object-centric — and we state where it does not work.

Tianjian Zhou[1,†] · **Jie Jiang**[1,†] · **Yishan Li**[1,*] · **Yifei Zhang**[2,*]
NeurIPS 2026 · Main Track · grant 72671303
[†] equal contribution · [*] corresponding author
[1] National University of Defense Technology · [2] Northwestern Polytechnical University

liyishan@nudt.edu.cn · yifeiacc@gmail.com
Code & configs github.com/Jaywalk18/DAGA



CODE WECHAT