JetViT: Efficient High-Resolution Vision Transformer with Post-Training Attention Search

Supplementary Material

Dongyun Zou, Zhuoyang Zhang, Junyu Chen, Wenkun He, Qinhe Peng, Hanrong Ye, Yao Lu, Hongxu Yin, Yu Wang, Song Han, Han Cai
MIT

A. Training Details

To efficiently conduct Post-Training Attention Search and minimize computational overhead, we primarily employ a **Progressive Resolution Scaling** strategy. This allows the model to adapt inherited weights and learn structural patterns at lower resolutions with reduced cost, before adapting to high-resolution details. Throughout the training process, we utilize feature distillation to align the student model with the teacher.

A.1. General Training Strategy

Progressive Resolution Scaling. Training Vision Transformers at high resolutions is computationally expensive. To address this, we initiate the training process at a lower resolution to quickly stabilize the inherited weights. Subsequently, we increase the resolution to capture fine-grained details required for dense prediction tasks. Throughout all stages, we maintain a constant batch size of 1024.

Feature Distillation Configuration. We align the student's features with the teacher's by minimizing the Mean Squared Error (MSE) loss. The application of distillation depends on the downstream task:

- **Semantic Segmentation:** Since we use a linear probing head on the final feature map, we apply MSE loss only between the last layer's output of the student and the teacher.
- **Monocular Depth Estimation:** As this task relies on a DPT head that fuses multi-scale features, we apply MSE loss to the four intermediate feature maps involved in the dense prediction process.

A.2. DINOv3 Post-Training Attention Search Training Pipeline

For the DINOv3-based model, we implement the progressive resolution scaling strategy, increasing resolution from 256^2 to 1024^2 across the search and fine-tuning stages.

Step 1: Linear and Window-Attention Search. Starting with the inherited weights, we first perform a low-resolution training phase to stabilize the supernetwork. We conduct feature distillation at a resolution of 256×256 for 3,000 steps. Following this, we increase the resolution to 512×512 and continue feature distillation for another 5,000 steps. Finally, we employ beam search on this supernetwork to identify the optimal combination of Linear-Attention and Window-Attention blocks.

Step 2: Full-Attention Placement Search. Based on the efficient architecture found in Step 1, we construct a new supernetwork containing both the searched efficient blocks and Full-Attention blocks. We perform feature distillation at 512×512 resolution for 5000 steps. Beam search is then applied to determine the critical locations for inserting Full-Attention blocks.

Final Fine-tuning. Once the final architecture is determined, we perform a final distillation stage at a high resolution of 1024×1024 for 10,000 steps. This step ensures the model extracts features of the same quality as teacher model on high resolution images.

A.3. DepthAnythingV2 Post-Training Attention Search Pipeline

DepthAnythingV2 models are trained on 518×518 resolution. We adapt our pipeline as follows:

Step 1: Linear and Window-Attention Search. We perform feature distillation for 5000 steps at 518×518 resolution. Then we employ beam search to find the optimal combination of linear-attention and window-attention.

Step 2: Full-Attention Placement Search. After converting the searched efficient architecture to supernetwork. We perform feature distillation for 5000 steps at 518×518 resolution. Beam search is then applied to find critical location for placing full-attention block.

Final Fine-tuning with Pseudo-Labels. After fixing the model structure, we fine-tune the model at a high resolution of 1022×1022 for 10,000 steps. Following DepthAnythingV2, we utilize pseudo-depth labels generated by the Depth Anything V2 Giant model on unlabeled real-world images. This allows Post-Training Attention Search to achieve competitive performance on high-resolution depth estimation benchmarks.

B. Computational Cost Analysis

A key advantage of Post-Training Attention Search is its low computational overhead. Since Post-Training Attention Search reuses pretrained weights and performs distillation on public datasets, the search cost is negligible compared to pretraining from scratch.

Model	Pre-train	JetViT	Ratio
DINOv3-7B	61,440 (H100)	**900 (H100)**	**1/68**
DepthAnyV2-G	—	**280 (A100)**	—

Table 1. **Computational cost comparison** (GPU hours). JetViT requires only $\sim$1/68 of the pretraining cost for DINOv3-7B. For DepthAnythingV2-Giant, our full pipeline requires only 280 A100 GPU hours.

C. Generalization to More Tasks

To further validate the generalizability of Post-Training Attention Search, we evaluate JetViT on object detection and image/video classification in addition to the segmentation and depth estimation results reported in the main paper.

C.1. Object Detection

We apply Post-Training Attention Search to the YOLOS [2] model and evaluate on COCO object detection. Latency is measured on an NVIDIA H100 GPU at 2K resolution following the same protocol as the main paper.

Model	mAP	Latency (ms)
YOLOS-Base [2]	42.0	51.1
JetViT-YOLOS-Base	42.0	**31.7**

Table 2. **Object detection on COCO.** JetViT achieves the same mAP as YOLOS-Base with 1.6$\times$ lower latency.

C.2. Image and Video Classification

We evaluate JetViT via linear probing on ImageNet [1] (image classification) and UCF101 (video classification), using DINO-Base as the teacher model.

Model	ImageNet (%)	UCF101 (%)
DINO-Base	83.7	91.6
JetViT-DINO-Base	83.7	91.3

Table 3. **Linear probing results on ImageNet and UCF101.** JetViT preserves the teacher's representational quality on both image and video classification.

D. Efficiency on Edge Devices

Beyond server-grade H100 GPUs, we evaluate JetViT on edge hardware to demonstrate practical deployment benefits. Latency is measured at 2K resolution on an NVIDIA RTX 3090 and a Jetson GB10.

	RTX 3090 (ms)		Jetson GB10 (ms)	
Model	Original	JetViT	Original	JetViT
DINOv3-L	670	**273**	482	**312**
DepthAnythingV2-L	1135	**472**	803	**529**

Table 4. **Latency comparison on edge devices at 2K resolution.** JetViT achieves consistent speedups of up to 2.4$\times$ on both consumer GPU (RTX 3090) and embedded GPU (Jetson GB10) platforms.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[2] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021. 2