ICML
International Conference
On Machine Learning

Attentive Multi-Layer Fusion for Vision Transformers

Laure Ciernik[*1,2,3], **Marco Morik**[*1,3], Lukas Thede[4,5,6,7], Luca Eyring[6,7,8], Shinichi Nakajima[1,3,9], Zeynep Akata[6,7,8], Lukas Muttenthaler[6,7,8,10]

** Equal contribution*

1 2 3 4 5

6 7 8 9 10

Does the final layer capture all task-relevant information in a ViT?

Head2Toe (Evci et al., 2022)

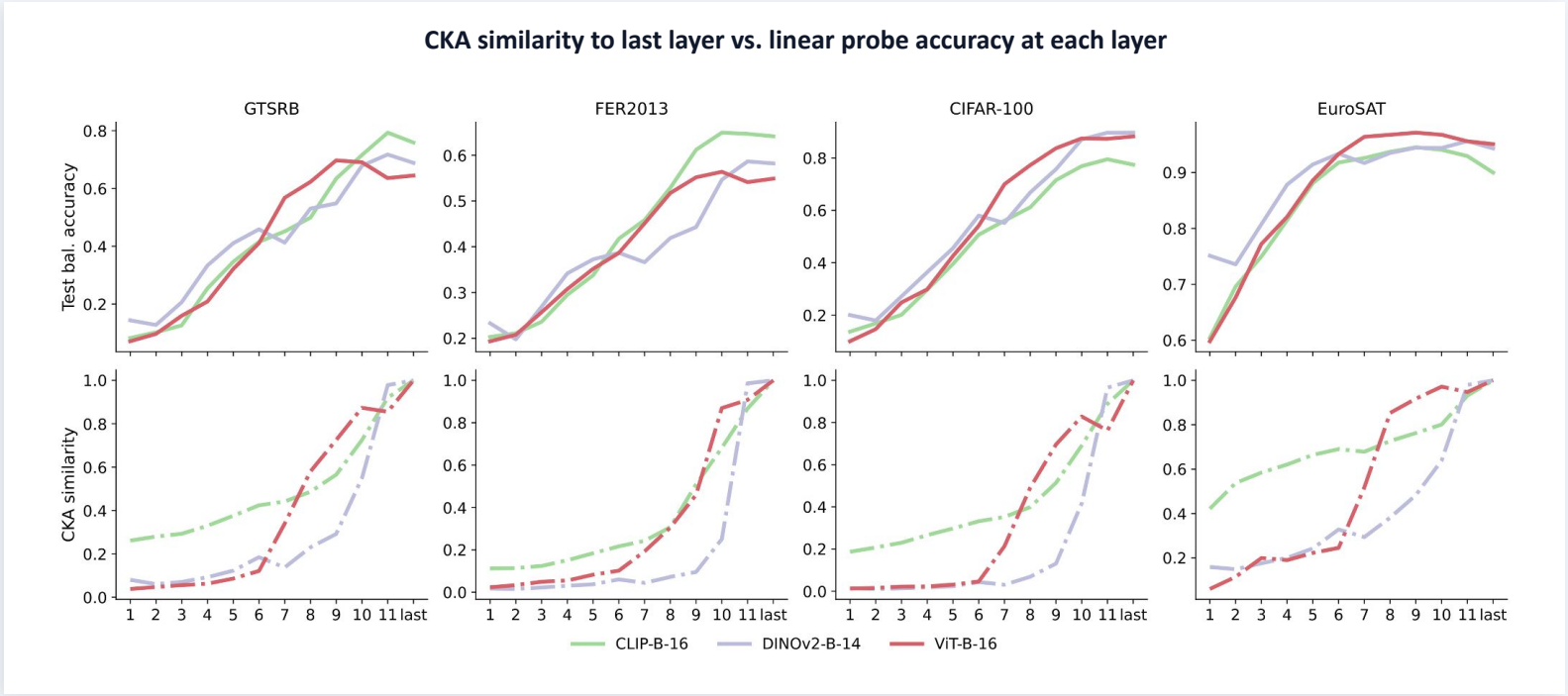
Trains a linear head connected to every intermediate node in the network and outperforms last-layer-only probing.

Visual Query Tuning (Tu et al., 2023)

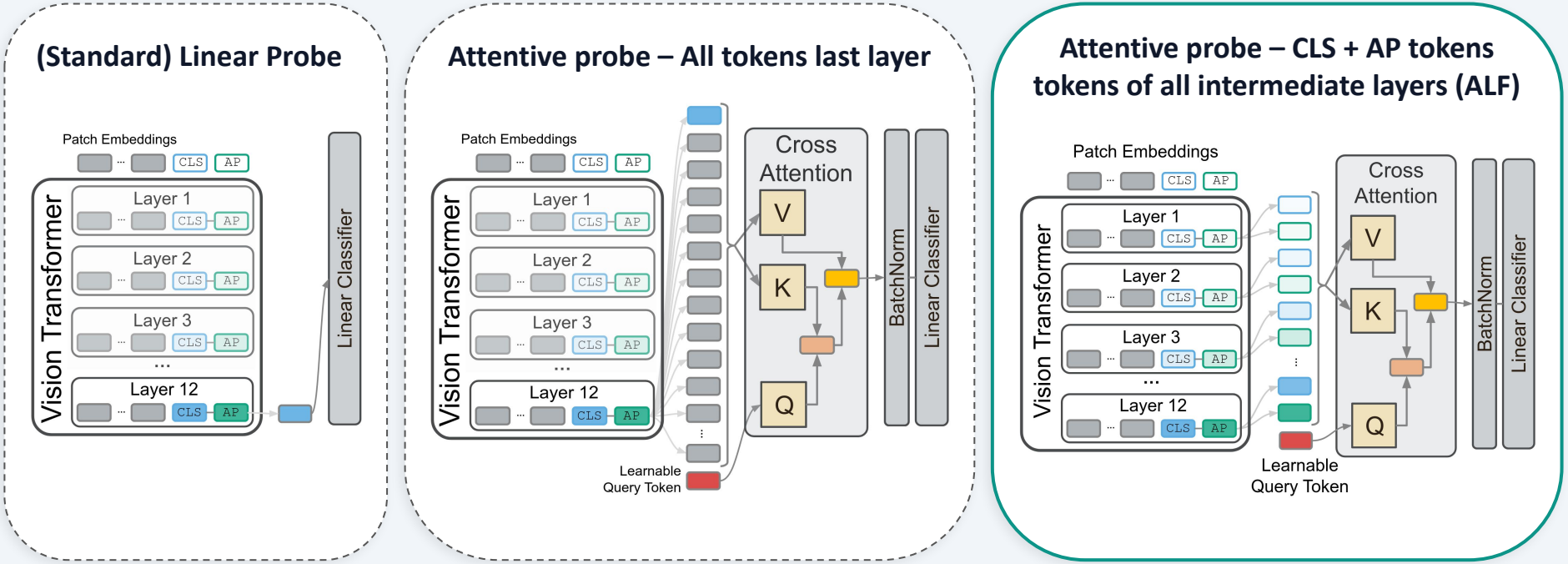
Intermediate representations improve parameter- and memory-efficient transfer via learnable query tokens.

DINOv2 (Oquab et al., 2024)

Concatenating CLS tokens from four last layers surpasses single-layer probing



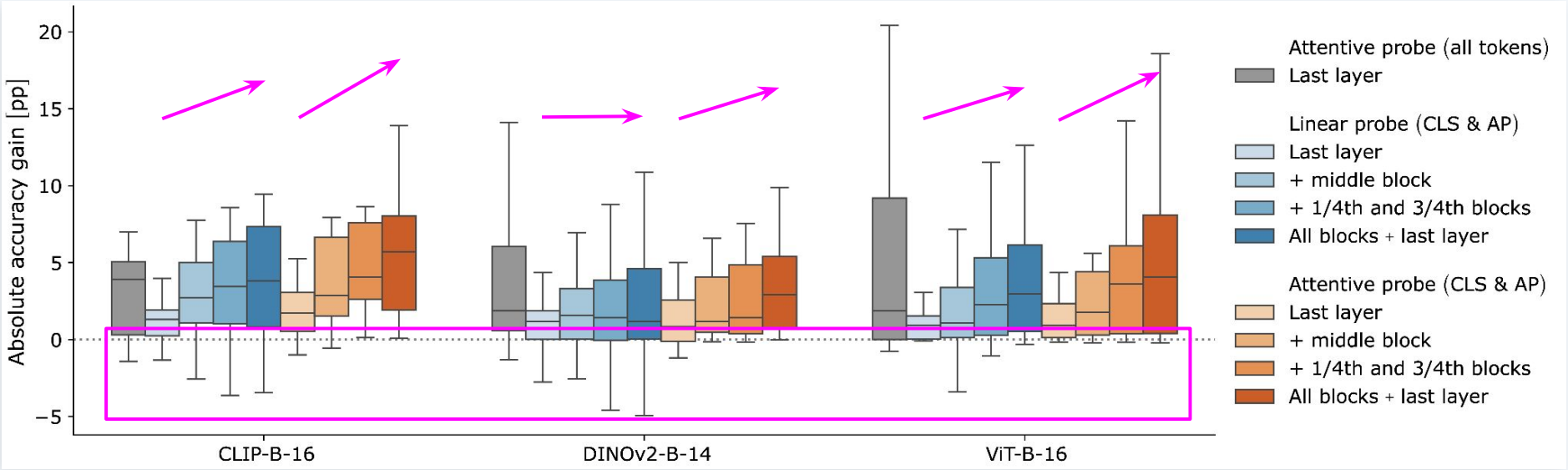
Attentive multi-layer fusion and baselines



Analysis across

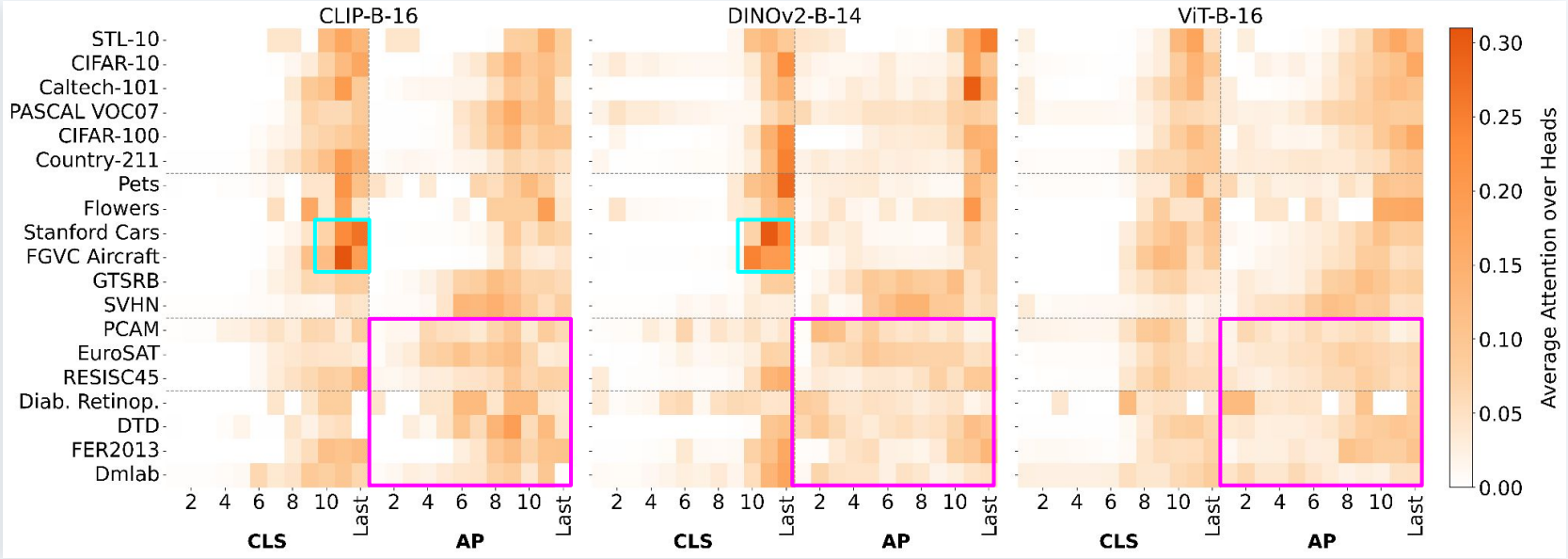
- **20 datasets (Natural multi & single domain, special and structured)**
- **9 pretrained ViTs of different capacities and trained with different objectives**

Attentive Layer Fusion Yields Robust Gains Over All Baselines



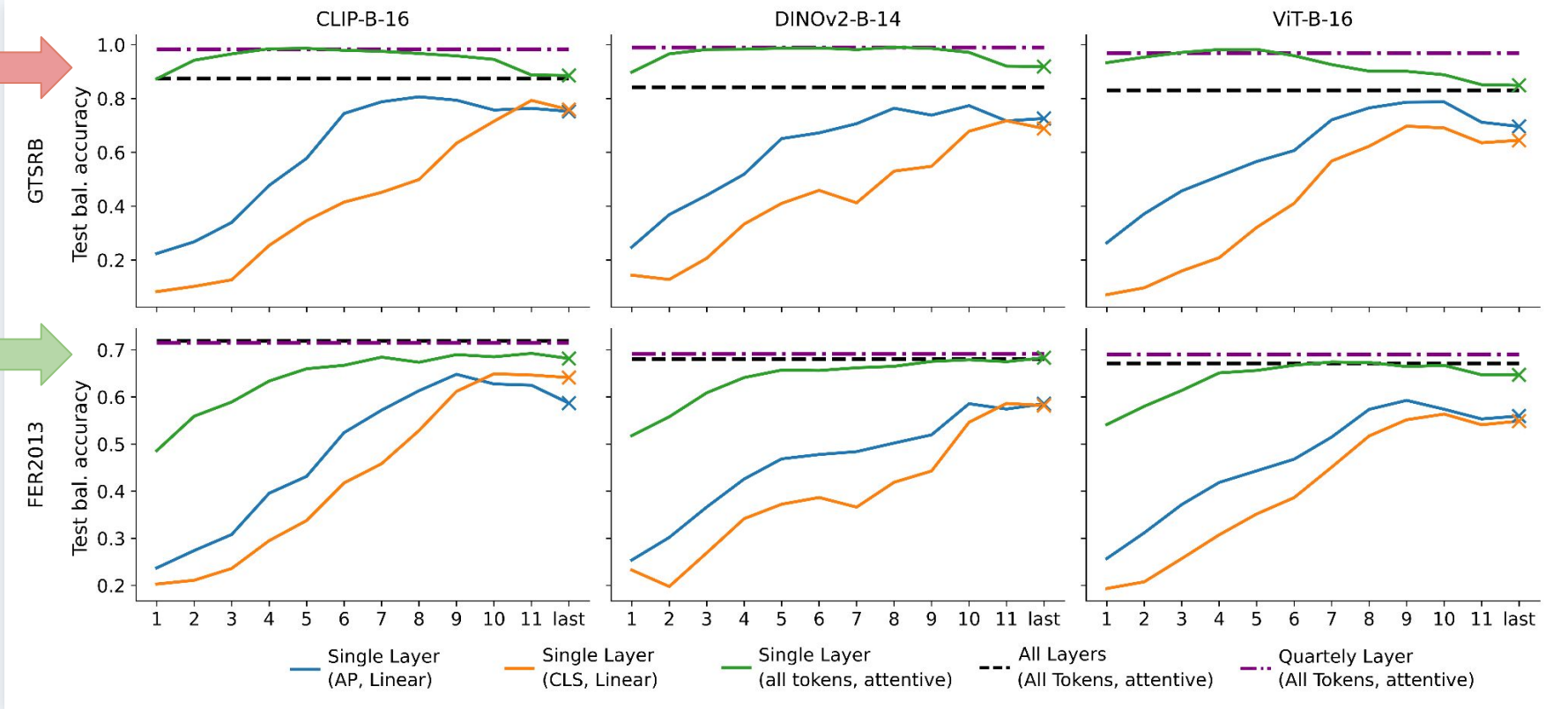
Additionally, it scales with model size, thus larger models still leave substantial information outside the final CLS token.

Adaptive layer selection to task-dependencies



Out-of-domain tasks (satellite, medical) benefit most; attention heatmaps shift to intermediate AP tokens for EuroSAT, FER2013 (Fig. 4, paper).

Spatial and hierarchical information axes are orthogonal



Takeaways

01

Intermediate layers are not redundant

CKA similarity to the last layer is low for early layers, yet their accuracy is comparable or higher — evidence of complementary, non-redundant information.

02

Fusion method is as important as layer selection

Naive concatenation is unstable (+high variance). Attentive-Probing adaptively re-weights layers per task, leading to improved performance

03

Spatial and hierarchical axes are orthogonal

ALF fuses across depth; patch-token attention-probing fuses across space. Both are needed for peak performance on fine-grained tasks.

04

More efficient than finetuning

36× faster in training than fine-tuning, <5% of backbone parameters, consistent gains across CLIP / DINOv2 / ViT / MAE at all model scales.

Thank you!

Visit us at our poster Wed, Jul 8
10:30 am to 12:15 pm KST
Coex: HALL A 🎉