REVEALING TASK-DEPENDENT LAYER RELEVANCE VIA ATTENTIVE MULTI-LAYER FUSION

Marco Morik[*,†]
Machine Learning Group
TU Berlin
BIFOLD [‡]

Laure Ciernik[*,†]
Machine Learning Group
TU Berlin
Hector Fellow Academy
ELLIS

Lukas Thede
University of Tübingen
Tübingen AI Center
Helmholtz Munich
MCML [§]

Luca Eyring
Helmholtz Munich
TU Munich
MCML [§]

Shinichi Nakajima
Machine Learning Group
TU Berlin
BIFOLD [‡]
RIKEN AIP

Zeynep Akata
Helmholtz Munich
TU Munich
MCML [§]

Lukas Muttenthaler [†]
Aignostics GmbH
Helmholtz Munich
TU Munich
MCML [§]

ABSTRACT

Efficiently adapting large-scale foundation models to downstream tasks is a central challenge in modern deep learning. While linear probing is a standard and computationally efficient method, it typically operates exclusively on the final layer's representation. In this work, we present experimental evidence that this approach discards crucial task-relevant information distributed across other layers of the network. To investigate this, we introduce Attentive Layer Fusion (ALF), a probing mechanism that dynamically fuses representations from all layers of Vision Transformers. Acting as a an investigation tool, ALF reveals that optimal representation depth is highly task-dependent: while tasks similar to the pre-training domain rely on the final layer, specialized domains (e.g., medical, satellite) benefit significantly from intermediate layers. Furthermore, by analyzing representational similarities, we show that intermediate layers often achieve high downstream performance despite having low similarity to the final layer, indicating they encode distinct, complementary features. Across 19 diverse datasets and 9 foundation models, our hierarchical approach achieves consistent gains, offering a new lens into how foundation models organize information.

1 INTRODUCTION

Foundation models (Radford et al., 2021; Oquab et al., 2024) are trained to build hierarchical abstractions of their input. It is generally understood that early layers capture low-level structural cues (edges, textures), while later layers encode high-level semantic concepts aligned with the pre-training objective (Raghu et al., 2021; Dorszewski et al., 2025). However, linear probing, the standard methodology for adapting these models to downstream tasks, typically uses only the final layer's `CLS` token.

This design implicitly assumes that the final layer is a sufficient statistic for all downstream tasks. While recent work has begun to challenge this assumption by concatenating late layers (Oquab et al., 2024) or using features across the network's hierarchy (Tu et al., 2023; Wu et al., 2024; Bolya et al., 2025), a systematic understanding about the specific location of task-relevant information within a model is missing. We hypothesize that as a model progresses through its depth, it compresses information to satisfy its pre-training objective, potentially "abstracting away" structural or textural details.

[*]Equal contribution.

[†]Correspondence to *m.morik@tu-berlin.de*,*ciernik@tu-berlin.de*, or *lukas.muttenthaler@tu-berlin.de*.
[‡]Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
[§]Munich Center for Machine Learning, Munich, Germany.

In this work, we propose **Attentive Layer Fusion (ALF)** as a probe to analyze the distribution of information within Vision Transformers (ViTs). ALF uses an attention probe Chen et al. (2024) that attends to summary tokens (both average-pooled `AP` and `CLS`) from *all* layers simultaneously, automatically discovering the most relevant abstraction level for a given task. Our experiments across 19 datasets reveal distinct empirical regularities. We find that intermediate layers are often superior to the final layer for tasks that differ from the pre-training domain. We use Centered Kernel Alignment (CKA)(Kornblith et al., 2019a) to show that layers with low similarity to the final output can still drive high performance. Additionally, attention heatmaps extracted from our probe reveal dataset-dependent patterns. For natural images, the model relies more on later layers, while for structured or specialized datasets, the intermediate layers appear more crucial. These findings suggest that future adaptation methods must be depth-aware, using the entire backbone as an input rather than just the final output tokens.

2 METHOD

To probe the hierarchy, we extract representations from all $\mathcal{L}$ layers of a frozen ViT. For each layer ℓ, we extract two complementary summary tokens: the global classification token $\boldsymbol{h}^{(\ell)}_{\texttt{[CLS]}}$ and the spatial average $\boldsymbol{h}^{(\ell)}_{\texttt{[AP]}}$ (Average Pool).

We include $\boldsymbol{h}^{(\ell)}_{\texttt{[AP]}}$ because the `[CLS]` token in early layers is often not fully contextualized, whereas spatial averaging captures low-level statistics and texture information that may be preserved throughout the hierarchy. We stack these to form a memory bank of hierarchical features $\boldsymbol{H}_{\mathcal{L}} \in \mathbb{R}^{2L \times d}$.

We employ a multi-head cross-attention mechanism to fuse these features Chen et al. (2024). A learnable query vector $\boldsymbol{Q}$ (acting as a "task prototype") attends to the layer representations. For each head m, we compute:

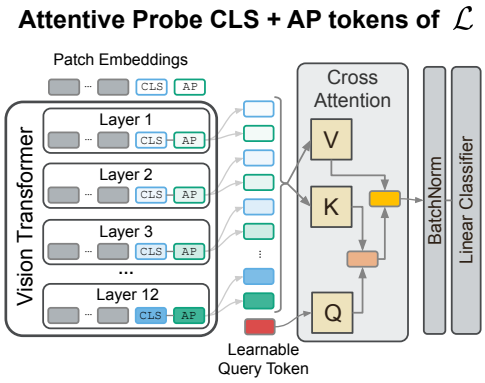

Figure 1: **Attentive Layer Fusion (ALF).** By treating layers as a sequence to be queried, ALF automatically discovers the most relevant abstraction level for a given task.

$$\text{Attention}(\boldsymbol{Q}^{(m)}, \boldsymbol{H}_{\mathcal{L}}) = \text{softmax}\left(\frac{\boldsymbol{Q}^{(m)}(\boldsymbol{H}_{\mathcal{L}}\boldsymbol{W}_k^{(m)})^\top}{\sqrt{d_h}}\right)(\boldsymbol{H}_{\mathcal{L}}\boldsymbol{W}_v^{(m)}) \tag{1}$$

The attention weights produced by the softmax provide a direct, interpretable measure of **layer relevance**. If the model assigns a high weight to Layer k, we infer that Layer k contains information critical for the task that is either lost or obfuscated in other layers. An illustration of our approach can be seen in Fig. 1.

3 EXPERIMENTS

We evaluate our hypothesis on 19 datasets from VTAB and the CLIP-benchmark, covering natural, specialized, and structured domains. We utilize 9 foundation models from three families: CLIP (image-text aligned), DINOv2 (self-supervised), and supervised ViTs, spanning Small, Base, and Large sizes. Full experimental details are provided in Section A and the code is available under `https://github.com/lciernik/attentive-layer-fusion`.

3.1 INTERMEDIATE LAYERS CONTAIN DISTINCT, NON-REDUNDANT FEATURES

To understand the nature of the information distributed across the hierarchy, we first examine the relationship between downstream performance and representational similarity. We train linear probes on each layer individually and compute the CKA similarity with an RBF kernel ($\sigma = 0.2$) (Kornblith et al., 2019a) between each intermediate layer and the final layer representation.

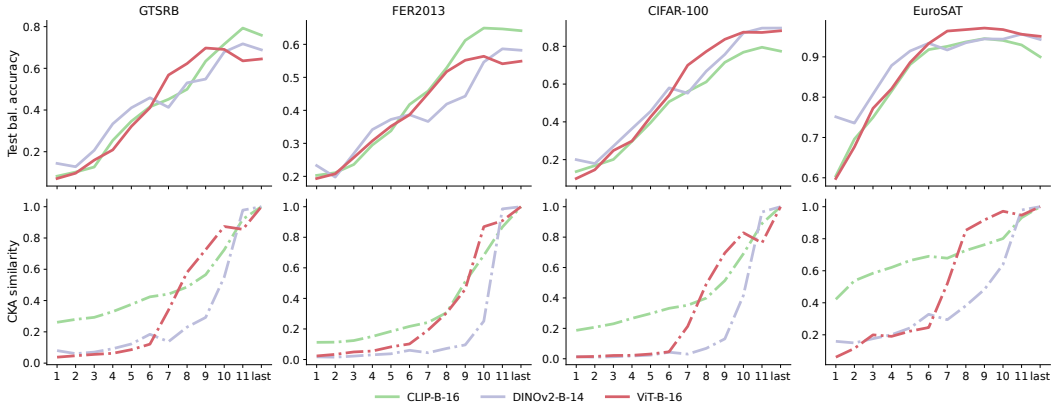

Figure 2: **Downstream performance vs. Representational Similarity. Top:** Balanced accuracy of linear probes trained on individual layers. **Bottom:** CKA similarity between Layer i and the Final Layer. On specialized tasks like GTSRB or EuroSAT, performance often peaks in middle layers despite these layers having very low similarity to the final representation. This indicates that vital task information is discarded in the final layers.

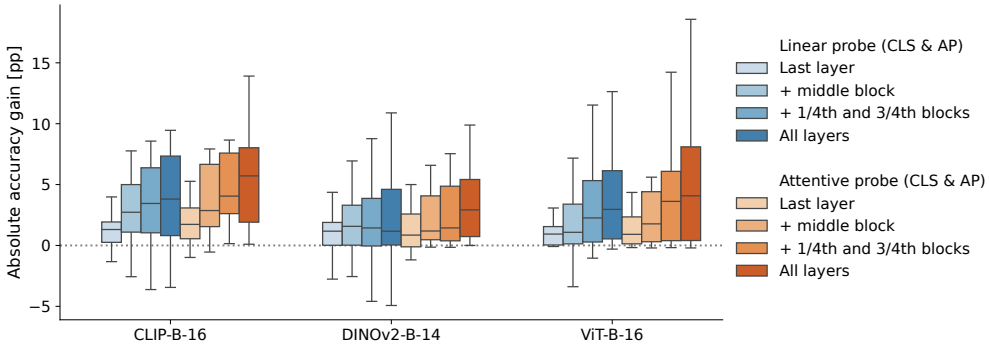

Figure 3: **Performance gain over standard Linear Probing.** Adding intermediate layers consistently improves performance across CLIP, DINOv2, and ViT backbones. ALF (orange) effectively selects relevant information, whereas naive linear concatenation (blue) suffers from high variance and instability.

Fig. 2 reveals a disconnect between semantic proximity and task utility. As shown in the bottom row, the CKA similarity to the final layer remains low throughout the network, rising sharply only in the final 2-3 layers. If the final layer were a comprehensive summary, we would expect accuracy to follow a similar trend. However, the top row shows that downstream accuracy often rises much earlier and, crucially, frequently peaks in the middle of the network.

This phenomenon is most visible on out-of-distribution tasks. For example, on EuroSAT (satellite imagery), the DINOv2 backbone achieves high accuracy around layers 6-8, where its CKA similarity to the final layer is negligible (< 0.3). As the model progresses to layer 12, similarity to the final state naturally maximizes, but downstream accuracy often stagnates. Two explanations could account for this behavior. Either the final layer is primarily a refinement of earlier representations, in which case intermediate layers add little beyond redundancy. Alternatively, the network reallocates information across depth, trading structural cues for more abstract semantics to satisfy its pre-training objective. In that case intermediate layers retain signals not preserved in the final representation.

3.2 Information is Distributed Hierarchically

We next quantify the benefit of accessing this distributed information. We measure the absolute accuracy gain relative to a standard linear probe on the final `[CLS]` token. To separate the benefit of information availability (access to the layers) from information selection (which layers to use), we compare two fusion strategies: naive Linear Concatenation (blue) and Attentive Layer Fusion (ALF, orange). For both, we vary the input scope: accessing only the last layer, adding the middle layer, adding quarterly intervals, and finally accessing the complete hierarchy.

As shown in Fig. 3, performance scales positively with the inclusion of intermediate layers. Merely adding the middle layer improves accuracy over the baseline, and accessing the full hierarchy yields the highest median gain.

Crucially, these results highlight the distinction between availability and accessibility. While naive linear concatenation of all layers improves performance on average, it suffers from high variance,

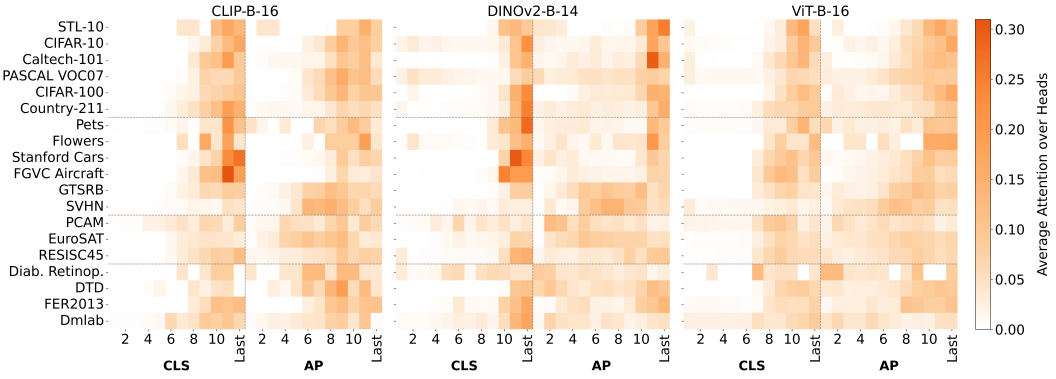

Figure 4: **Layer Relevance Heatmaps.** We visualize the learned attention weights for different datasets on Base models. **Left:** Natural image tasks (CIFAR, Pets) focus heavily on the final layers (11-12). **Right:** Out-of-distribution tasks (EuroSAT, textures, medical) shift attention to intermediate layers (5-9), indicating that the final layer has abstracted away necessary structural information.

leading to performance degradation on some tasks due to the curse of dimensionality. In contrast, ALF acts as a selective filter. By dynamically downweighting irrelevant layers, it achieves consistent gains (≈ 5.5 percentage points on average) without performance deficits. This confirms that while task-relevant information is distributed throughout the hierarchy, it must be carefully selected.

3.3 Domain Shift Dictates Layer Relevance

Finally, we visualize how ALF utilizes this distributed information. Fig. 4 plots the learned attention weights, acting effectively as a "depth-meter" for domain shift.

The heatmaps reveal two distinct behaviors depending on the downstream task. For **in-distribution (Natural Images)** datasets such as CIFAR-10, Oxford Pets, or Flowers102, the model attends primarily to the final layers, often prioritizing the `CLS` token. This validates that for tasks semantically aligned with the pre-training data, the standard assumption, that the final layer contains the optimal representation, holds true.

In contrast, for **out-of-distribution (Specialized/Structured)** datasets involving textures (DTD), satellite imagery (EuroSAT, Resisc45), or medical data (PCAM), the center of gravity shifts significantly towards intermediate layers and the spatial `AP` token. In the case of EuroSAT, the model explicitly retrieves information from the `AP` tokens of layers 6-9. This aligns directly with our earlier CKA analysis (Section 3.1), confirming that the model "reaches back" to retrieve structural features that were transformed or discarded in the final semantic abstraction. Overall, these heatmaps demonstrate that ALF serves not just as an adaptation method but also as an accessible diagnostic tool for analyzing the layer-wise utility of foundation models for any given task.

4 Conclusion and Discussion

In this paper, we presented Attentive Layer Fusion (ALF), a method that leverages the full depth of Vision Transformers to improve downstream adaptation. Beyond performance gains, our work serves as an investigation tool into the internal structure of foundation models, challenging the widespread view that the final layer is the only source of task-relevant information (Raghu et al., 2021; Kornblith et al., 2019b). We provide empirical evidence that the last layer is often an arbitrary truncation point. For specialized domains like medical or satellite imagery, the most valuable signals are effectively "left behind" in the middle of the network.

Crucially, we demonstrate that accessing this distributed information requires more than simple concatenation. While a naive linear combination of layers leads to instability and overfitting, our attentive mechanism acts as a selective filter, reliably identifying the optimal level of abstraction for each task. These observations align with similar findings in Language Models (Skean et al., 2025), suggesting a fundamental principle for foundation models across modalities: as models compress information to satisfy their pre-training objectives, they abstract away details that are crucial for specialized downstream tasks. Consequently, future research, whether in vision, language, or emerging biological models (Brixi et al., 2025), should view adaptation not as a simple mapping from the final output, but as a search for task-relevant information across all layers of a model.

REFERENCES

Yuri Sousa Aurelio, Gustavo Matheus de Almeida, Cristiano Leite de Castro, and Antonio Padua Braga. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Processing Letters*, 50(2), 2019. doi: 10.1007/s11063-018-09977-1. URL https://doi.org/10.1007/s11063-018-09977-1.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=INqBOmwIpG.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K Wang, Etowah Adams, Stephen A Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y Ng, Jaspreet Pannu, Christopher Re, Jonathan C Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Tom McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D Hsu, and Brian Hie. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1), 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 105(10), 2017. doi: 10.1109/JPROC.2017.2675998. URL https://doi.org/10.1109/JPROC.2017.2675998.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=va3zmBXPat.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. doi: 10.1109/CVPR.2009.5206848.

Teresa Dorszewski, Lenka Tětková, Robert Jenssen, Lars Kai Hansen, and Kristoffer Knutsen Wickstrøm. From colors to classes: Emergence of concepts in vision transformers. *arXiv preprint arXiv:2503.24071*, 2025.

Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection, 2015. URL https://kaggle.com/competitions/diabetic-retinopathy-detection. Kaggle.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010. doi: 10.1007/S11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 2006. doi: 10.1109/TPAMI.2006.79.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron C. Courville, Mehdi Mirza, Benjamin Hamner, William Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Tudor Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. volume 64, 2015. doi: 10.1016/J.NEUNET.2014.09.005. URL https://doi.org/10.1016/j.neunet.2014.09.005.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97, 2019a. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV) Workshops*, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Lukas Muttenthaler and Martin N. Hebart. Thingsvision: A python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics*, 15, 2021. URL https://www.frontiersin.org/article/10.3389/fninf.2021.679838.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. Imagenet-21k pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98f13708210194c475687be6106a3b84-Paper-round1.pdf.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/J.NEUNET.2012.02.016. URL https://doi.org/10.1016/j.neunet.2012.02.016.

Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 11071 of *Lecture Notes in Computer Science*, 2018. URL https://doi.org/10.1007/978-3-030-00934-2_24.

Zhi-Fan Wu, Chaojie Mao, Xue Wang, Jianwen Jiang, Yiliang Lv, and Rong Jin. Structured model probing: Empowering efficient transfer learning by structured regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2020. URL https://arxiv.org/abs/1910.04867.

A Implementation Details

Table 1: Overview of the 19 datasets used in our experiments including the size of both train and test set, number of classes, and the Class Imbalance Ratio (CIR) calculated by $\frac{N_{\text{Majority Class}}}{N_{\text{Minority Class}}}$.

Category	Dataset	Train Size	Test Size	Classes	CIR	Reference
Natural (MD)	STL-10	5 000	8 000	10	1	Coates et al. (2011)
	CIFAR-10	45 000	10 000	10	1.02	Krizhevsky (2009)
	Caltech-101	2 753	6 085	102	1.3	Fei-Fei et al. (2006)
	PASCAL VOC 2007	7 844	14 976	20	20.65	Everingham et al. (2010)
	CIFAR-100	45 000	10 000	100	1.06	Krizhevsky (2009)
	Country-211	31 650	21 100	211	1	Radford et al. (2021)
Natural (SD)	Pets	2 944	3 669	37	1.24	Parkhi et al. (2012)
	Flowers	1 020	6 149	102	1	Nilsback & Zisserman (2008)
	Stanford Cars	8 144	8 041	196	2.83	Krause et al. (2013)
	FGVC Aircraft	3 334	3 333	100	1.03	Maji et al. (2013)
	GTSRB	26 640	12 630	43	10	Stallkamp et al. (2012)
	SVHN	65 931	26 032	10	2.98	Netzer et al. (2011)
Specialized	PCAM	262 144	32 768	2	1	Veeling et al. (2018)
	EuroSAT	16 200	5 400	10	1.58	Helber et al. (2019)
	RESISC45	18 900	6 300	45	1.16	Cheng et al. (2017)
	Diabetic Retinopathy	35 126	42 670	5	36.45	Dugas et al. (2015)
Structured	DTD	1 880	1 880	47	1	Cimpoi et al. (2014)
	FER2013	28 709	7 178	7	16.55	Goodfellow et al. (2015)
	Dmlab	65 550	22 735	6	1.98	Zhai et al. (2020)

Feature Extraction & Preprocessing. We extract representations using `thingsvision` (Muttenthaler & Hebart, 2021). Images are resized to 256px and center-cropped to 224px. Extracted features are L2-normalized. We use a frozen backbone strategy, training only the fusion module and classifier. To handle class imbalance, we employ a weighted cross-entropy loss where class weights are inversely proportional to class frequency (Aurelio et al., 2019).

Training & Optimization. Models are trained for ≥ 40 epochs (ensuring ≥ 1000 steps) using AdamW with cosine annealing. We use a batch size of up to 2048. To prevent overfitting, we apply gradient clipping (norm 5.0), weight decay, and inject Gaussian noise ($\mathcal{N}(0, 0.05)$) to representations during training. Hyperparameters were selected via grid search on a validation split (80/20): Learning rates $\in \{10^{-1}, \ldots, 10^{-3}\}$, Dropout $\in \{0.0, 0.1, 0.3\}$, Weight decay $\in \{10^{-6}, \ldots, 1.0\}$.

Attention Mechanism. We base the attentive probe on the modle of Chen et al. (2024). For ALF, the number of attention heads M matches the number of fused representations (e.g., $M = 24$ for 12 layers of `CLS` + `AP`). Queries are initialized from $\mathcal{N}(0, 0.02)$. Code is based on Ciernik et al. (2025) and will be released.

Evaluation Metric To enable an intuitive comparison of performances across datasets, we report the absolute top-1 accuracy gain (in percentage points [pp]) of each method over the standard linear probe `CLS` baseline: $\text{Acc}_{\text{bal}}(\text{method}) - \text{Acc}_{\text{bal}}(\text{CLS}_{\text{linear}})$, which is positive if the method outperforms the baseline.

Model Details We evaluate three model families across Small, Base, and Large scales:

- **Supervised ViT:** ViT-S/16, ViT-B/16, and ViT-L/16 pretrained on ImageNet-21K and fine-tuned on ImageNet-1K (Deng et al., 2009; Ridnik et al., 2021).
- **Self-Supervised DINOv2:** ViT-S-14, ViT-B-14, and ViT-L-14 , pretrained on the LVD-142M dataset (Oquab et al., 2024).
- **Image-Text Alignment CLIP:** OpenCLIP models ViT-B-32, ViT-B-16, and ViT-L-14 (Cherti et al., 2023; Ilharco et al., 2021)) following the CLIP architecture and using its pretrained weights (Radford et al., 2021)).

Dataset Details An overview of all datasets used in this work is given in Tab. 1. Following VTAB (Zhai et al., 2020), the datasets are categorized by domain.