ON THE IMPORTANCE OF BACKBONE, PRETRAINING AND HYPERPARAMETER SELECTION FOR HIERARCHICAL FINE-GRAINED IMAGE RECOGNITION

Augusto Christian Surya (蘇立光)[1*], Edwin Arkel Rios (冉恩達)[2*],
Bo-Cheng Lai (賴伯承)[2], Min-Chun Hu (胡敏君)[1]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[2]Department of Electronics Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
E-mails: ss110000162@gapp.nthu.edu.tw, edwinarkel.rios@gmail.com,
bclai@nycu.edu.tw, anitahu@cs.nthu.edu.tw

ABSTRACT

Fine-grained image recognition (FGIR) is a task that requires models to distinguish visually similar subcategories within a broader class. FGIR tasks are naturally organized into hierarchical structures where coarse-level groupings provide valuable semantic context for fine-level classification. Despite this, most prior work in hierarchical FGIR evaluates only a limited set of backbone architectures, leaving the influence of backbone and pretraining largely underexplored. This gap is critical, as the backbone determines the quality of feature representations used across multiple classification heads in hierarchical settings. In this work, we systematically study 21 pretrained models, using convolutional and transformer-based backbones, evaluated under both fine-tuned and frozen backbone settings. We also evaluate 18 transferability metrics to examine their ability to predict performance. Our results show that backbone and hyperparameter selection heavily influence FGIR performance, with Discriminative models such as MoCov3 and SwaV on Resnet being a solid choice due to having desirable qualities, high mean accuracy and low standard deviation. Most of the metrics weakly correlates or inconsistent with transfer accuracy across different settings, highlighting an open challenge for future research.

Keywords: Fine-Grained Object Categorization, Hierarchical Classification, IPPR, CVGIP 2025.

1. INTRODUCTION

Fine-grained image recognition (FGIR) is a computer vision task to distinguish between visually similar subcategories within a broader class. While fine-grained categories are organized into taxonomic hierarchies, early works on FGIR approaches often treat categories as flat labels [1, 2], ignoring semantic relationships between them. This limited perspective prevents models from fully leveraging hierarchical relationships, leading to suboptimal performance. [3]

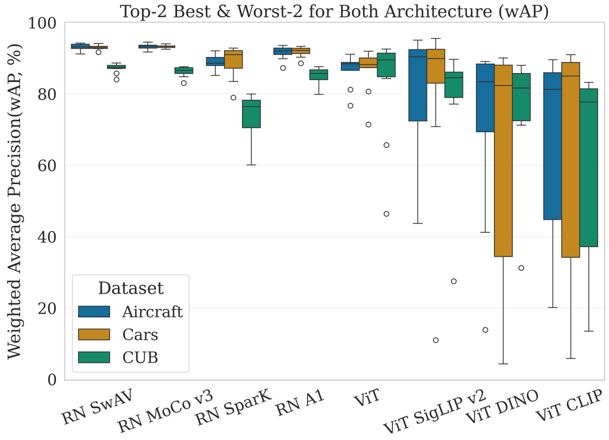

Fig. 1: Comparison of backbone performance. We visualize the Top-2 and Bottom-2 performing ResNet and ViT models respectively, ranked by Weighted Average Precision.

Hierarchical FGIR addresses this limitation by leveraging semantic relationships across different levels of granularity, using information from coarse categories to enhance fine-grained classification performance. Several approaches have explored this direction, such as granularity specific classifiers [4] and hierarchically discriminative loss functions [3]. While these directions have shown promise, they overlook a more foundational component: the backbone itself.

Although prior work in standard FGIR has shown that different backbones can lead to a significant impact on recognition performance [5], the role of backbone choice in hierarchical FGIR remains underexplored. Most hierarchical FGIR methods continue to rely on a narrow set of backbones [3, 4, 6–9], making it unclear whether backbone differences affect the performance in taxonomic hierarchies. This raises an important question: **To what extent does the back-**

[*]Equal contribution.

bone influence performance in hierarchical FGIR?

In this work, we conduct **the first study on backbone selection for hierarchical FGIR**. We benchmark a wide range of convolutional and transformer-based models across multiple datasets and hyperparameters, showing that backbone choice alone can yield drastically different outcomes, from state-of-the-art performance(91% wAP on CUB for base ViT), to poor results (80% wAP on CUB for CLIP), as observed in fig. 1. To understand this gap, we evaluate 7 metrics from prior work [10, 11] and introduce 11 additional metrics. While effective in standard classification, many of these metrics fail to generalize to hierarchical FGIR. This **underscores a disconnect between general transferability and hierarchical recognition**, and highlights the need for better predictors, especially given the cost of exhaustive model and hyperparameter search [12, 13].

Our study offers three main contributions:

1. We conduct an extensive study on the role of 11 CNN-based and 10 ViT-based backbone and pretraining in determining hierarchical FGIR performance. The right choice of backbone can **change up to 11% wAP** from worst to best.
2. We study the impact of hyperparameters across different models. The right choice of hyperparameter can **change up to 80% wAP** from worst to best.
3. Our analysis of 18 predictive highlights the challenge of predicting the performance hierarchical FGIR.

2. RELATED WORKS

2.1. Fine-Grained Image Recognition(FGIR)

Early FGIR methods relied on additional human supervision, such as bounding boxes or part annotations, to improve discriminative localization [1]. While effective, these methods are difficult to scale due to the cost of annotations. Later approaches shifted toward using image-level labels combined with attention mechanisms to discover informative regions. For example, MA-CNN [2] automatically identified key parts without requiring part-level labels. However, these models largely ignored class hierarchies, limiting their ability to capture semantic structure and generalize across categories.

2.2. Hierarchical Fine-Grained Image Recognition

To address these limitations, recent methods have incorporated multi-granularity structures into FGIR commonly referred to as *Hierarchical FGIR*. One of the first works, Multi-Granularity Descriptor (MGD) [6], used weakly supervised taxonomic descriptors. This was followed by methods like Hierarchical Semantic Embedding (HSE) [3] and granularity-specific classifiers [4], which introduced hierarchy-aware loss functions and multi-level prediction heads. While these approaches improved performance by leveraging taxonomic structure, they often still struggle to disambiguate classes within the same coarse category and current hierarchical FGIR methods predominantly rely on a limited subset of backbone architectures.

2.3. Backbone effect on FGIR

Recent work has explored the impact of backbone architectures on transfer learning and FGIR. Kornblith et al. [10] investigated whether ImageNet accuracy correlates with downstream performance on fine-grained tasks, concluding that it generalizes well. Nayman et al. [11] extended Kornblith et al.'s work by introducing alternative metrics to better predict transferability across tasks. However, both studies focus on general vision benchmarks and do not analyze hierarchical recognition explicitly, leaving open questions about how backbone design affects hierarchical FGIR.

3. EXPERIMENT SETUP

3.1. Architecture Overview

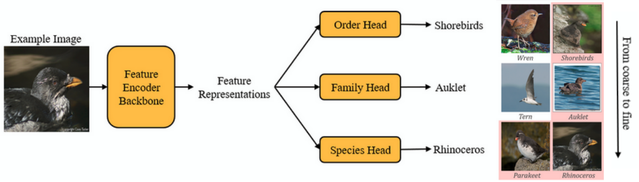

Fig. 2: Overall Architecture of the pipeline for this experiment

Our architecture is based on the hierarchical pipeline of Chang et al. [4]. An input image is passed through a feature encoder (backbone), whose output is shared across classification heads at multiple levels of semantic granularity (Fig. 2). We use 2 or 3-level taxonomy, where each level has its own classifier head to predict coarse-to-fine labels.

The backbone is central to the architecture: its representational quality directly impacts the accuracy of all hierarchical predictions. By varying the backbone while keeping the rest of the pipeline fixed, we isolate and evaluate its contribution to hierarchical FGIR performance.

3.2. Models

We evaluate 21 pretrained models across two backbone types, **Transformer (ViT-B)** and **ResNet50**, covering a broad range of training methods, including fully supervised (FSL), semi-supervised (SemiSL), and self-supervised learning (SSL). Here are the exact models used:

- **ViT-B models**:
  - **FSL**: ViT [14], DeiT [15], DeiT3 (IN1K, IN21K) [16], MIIL [17]
  - **SSL (Discriminative)**: MoCoV3 [18], DINO [19], CLIP-LAION5B [20], SigLIP-v2 [21]
  - **SSL (Generative)**: MAE [22]
- **ResNet50 models**:
  - **FSL**: Torchv1 [23], Torchv2 [23], Gluon [24], A1

[25], MIIL [17]

  - **SemiSL**: IG1B [26], YFCC100M [26]
  - **SSL (Discriminative)**: MoCoV3 [18], SupCon [27], SwAV [28]
  - **SSL (Generative)**: SparK [29]

3.3. Datasets

We evaluate the models on three widely used FGIR benchmarks: **FGVC Aircraft [30], CUB-200-2011 [31], and Stanford Cars [32]**. Detailed specifications of these datasets are provided in Table 1.

Dataset	Classes- L1	Classes- L2	Classes- L3	Split (Train)	Split (Test)
FGVC Aircraft	30	70	100	6667	3333
CUB-200-2011	13	38	200	5994	5794
Stanford Cars	-	9	196	8144	8041

Table 1: Datasets examined for FGIR transfer.

The hierarchical label taxonomies for these datasets, as described by Chang et al. [4], are constructed by tracing parent nodes (superclasses) from Wikipedia pages. This process ensures that the hierarchical structure reflects natural semantic relationships between the classes.

- **FGVC Aircraft** consists of 10,000 images representing 100 model variants of aircraft. The dataset follows a three-level label hierarchy: 30 manufacturers (makers), 70 families, and 100 plane models.
- **CUB-200-2011** contains 11,788 images from 200 bird species. It follows a similar three-level hierarchy, with 13 orders, 38 families, and 200 species.
- **Stanford Cars** includes 16,185 images of cars, categorized into 196 model variants. To form a two-level label hierarchy, an additional 9 car types are included.

3.4. Evaluation Targets

To assess model effectiveness in hierarchical FGIR, we measure several evaluation metrics capturing accuracy, class balance, and robustness:

- **Per-level accuracy**. The Top-1 classification accuracy of each level in the hierarchy. There are **3 hierarchical levels** at most, these are denoted by **L3, L2, L1, from coarsest to finest**.
- **Weighted Average Precision (wAP).** to address class imbalance, wAP weights per-class accuracy by sample count wAP is defined as:

$$wAP = \frac{1}{N}\sum_{k=1}^{K} n_k \cdot \mathrm{Acc}_k$$

  where K is the number of classes, n_k is the number of samples in class k, Acc_k is the accuracy for class k, and $N = \sum_{k=1}^{K} n_k$ is the total number of samples.

3.5. Representation-Based Metrics

To investigate which model properties are predictive of downstream hierarchical FGIR performance, we evaluate a set of 18 representation based metrics by computing their **Spearman rank correlation** (ρ) with multiple evaluation targets. These metrics capture different aspects of a model's features or pretraining characteristics, they include:

- **7 metrics from prior work**s [10, 11]: *ImageNet1k Accuracy, CKA Average, MSC, Intra-Class, Inter-Class, CIS (Clustering), CIS (Spectral)*
- and **11 additional metrics**: *CKA First Layer, CKA Last Layer, CKA Low Layers, CKA Mid Layers, CKA High Layers, Normalized Distance, L2 Norm, CIS (CKA First Layer), CIS (CKA Last Layer), CIS (Distance First Layer), CIS (Distance Last Layer).*

This setup allows us to quantify how well each metric aligns with model performance in a fine-grained hierarchical setting. These metrics are described in detail below:

- **ImageNet1k Top-1 Accuracy** A widely used benchmark for pretrained model effectiveness [10, 11]. We adapt this approach to investigate whether ImageNet1k top-1 accuracy can serve as an effective predictor for transfer accuracy in the context of hierarchical FGIR.
- **CKA (Central Kernel Alignment)** [33] CKA measures representational similarity between network layers by computing the normalized dot product between Gram matrices X^TX and Y^TY of feature activations X and Y. This metric has been shown to correlate with model transferability by [11]. CKA is defined below:

$$\mathrm{CKA}_{\mathrm{linear}}(X,Y) = \frac{\mathrm{vec}(X^TX)\cdot\mathrm{vec}(Y^TY)}{\|X^TX\|_F\|Y^TY\|_F}$$

  We compute CKA across all `bn2` layers in ResNet and `norm2` layers in ViT, further divided into three equal layer segments early, middle, and late corresponding to CKA (low, mid, high). Each segment comprises one-third of the total layers used (e.g., if 12 layers are selected, each group contains 4 layers).
- **Intra-class Variance & Inter-class Separation** [34] These 2 metrics quantify feature compactness within classes and separability between classes using embeddings from the penultimate layer respectively.
  Evaluation of both intra-class variance and inter-class separation is defined below [33]:

$$V_{\mathrm{intra}} = \sum_{k=1}^{K}\sum_{m=1}^{N_k}\sum_{n=1}^{N_k}\frac{1-\cos(\mathbf{x}_{k,m},\mathbf{x}_{k,n})}{KN_k^2} \quad (1)$$

$$S_{\mathrm{inter}} = \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{m=1}^{N_k}\sum_{n=1}^{N_j}\frac{1-\cos(\mathbf{x}_{k,m},\mathbf{x}_{j,n})}{K^2N_kN_j} \quad (2)$$

  Here, $\mathbf{x}_{k,m}$ represents the embedding of the m-th sample in class $k \in \{1, \ldots, K\}$, and N_k denotes the number of samples in class k.

- **Mean Silhouette Coefficient (MSC)** [35] MSC is used to measure the quality of clustering, capturing both intra-class variance and inter-class distance. The Silhouette Coefficient (SCm) is the relationship between intra-class distances (Vm) and nearest-class distances (Sm) as follows:

$$SC_m = \frac{s_m - v_m}{\max(s_m - v_m)}$$

$$v_m = \sum_{n=1}^{N_k} \frac{1 - \cos(x_{k,m}, x_{k,n})}{N_k - 1}$$

$$s_m = \min_{j \in \{1,\ldots,K\} \setminus \{k\}} \sum_{n=1}^{N_j} \frac{1 - \cos(x_{k,m}, x_{j,n})}{N_j}$$

Mean Silhouette Coefficient (MSC) is calculated as:

$$MSC = \frac{\sum_{k=1}^{K} \sum_{m=1}^{N_k} SC_m}{\sum_{k=1}^{K} N_k}$$

$x_{k,m}$ represents the embedding of the m-th sample in class $k \in \{1, \ldots, K\}$, and N_k denotes the number of samples in class k. MSC values range from -1 to 1, where higher values indicate better-defined clusters and greater separability between classes. The MSC is computed using feature embeddings extracted from the penultimate layer.

- **Calibrated Imagenet Score (CIS)** Introduced by [11], This metric combines ImageNet accuracy with feature diversity to better predict transferability.
CIS is defined below.

$$\text{CIS} = \text{ImageNet Accuracy} \times \text{Feature Diversity}$$

There are 2 feature diversities:
Clustering diversity: Average area under the agglomerative clustering ratio curve. Defined as

$$D_C = \frac{1}{L} \sum_{l=1}^{L} D_C(W^{(l)})$$

$$D_C(W^{(l)}) = \int_0^1 C_r(W^{(l)})\, d\tau$$

where $C_\tau(W^{(l)})$ is the cluster ratio at threshold r.

Spectral diversity: A measure of feature distribution using singular value decomposition (SVD).
The Spectral diversity of a layer l is defined as

$$D_S = \frac{1}{L} \sum_{l=1}^{L} D_S(W^{(l)}), \quad \text{where} \quad D_S(W^{(l)}) = \frac{\sum_{i=0}^{K} w_i}{\sum_{i=0}^{\infty} w_i},$$

$D_S(W^{(l)})$ are the singular values of the feature covariance matrix $W^T W$
Inspired by the CIS formulation, we further explore whether diversity can be captured by other structural similarity metrics. We define four new variants: **CIS (CKA First Layer)**, **CIS (CKA Last Layer)**, **CIS (Distance First Layer)**, **CIS (Distance Last Layer)**.

3.6. Training Settings

We evaluate two different training settings: **Fine-Tuning** and **Frozen Backbone**, to understand the impact of parameter adaptation during the training process.

- **Fine-Tuning (FT):** All backbone parameters are updated on the target dataset.
- **Frozen backbone (FZ):** Only classifier heads are trained, the backbone remains fixed.

3.7. Hyperparameters and Optimization

We conducted our experiments using the following set of hyperparameters and training configurations:

- SGD: {0.3, 0.1, 0.03, 0.01, 0.003}
- AdamW: {0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001, 0.000005}

Weight decay is set to 0 (SGD) and 0.05 (AdamW).

We trained the models for a total of 200 epochs with batch size of 8 images. We used Cosine Anneal Scheduler and used Mixed Precision Training. Images were resized to a square shape of 550x550 pixels. During training, random cropping was applied to generate patches of size 448x448 pixels, while during inference, the center area was cropped to the same size (Random Cropping), a horizontal flip was applied during training to augment the dataset.

4. RESULTS

4.1. Impact of Backbone and Pretraining

4.1.1. Choice of backbone and hyperparameter do matter

The choice of backbone and hyperparameter configuration has a significant impact on hierarchical FGIR performance. As shown in Table 2, applying the pipeline from Chang et al. [4] , and **through careful selection of backbones and hyperparameters**, we were able to achieve **state-of-the-art accuracy on both ResNet and ViT models** even **with older architecture**. This underscores the importance of backbone selection and hyperparameter tuning in optimizing hierarchical FGIR performance.

4.1.2. Distinct differences between ResNet and ViT models

The performance of both ResNet and ViT models show distinct patterns. In general, **ResNet models are more robust than ViT models regardless of pretrainings and hyperparameters**, seen in fig. 3. The **gap between worst fine-tuned and best frozen is small**, especially for ViT mod-

	CUB-200-2011 (CUB)				FGVC Aircraft				Stanford Cars		
Models	L3	L2	L1	wAP	L3	L2	L1	wAP	L2	L1	wAP
HMC-LMLP [7]	0.985	0.942	0.796	0.828	0.971	0.944	0.903	0.928	0.97	0.877	0.881
HMCN [36]	0.973	0.932	0.798	0.827	0.961	0.926	0.872	0.904	0.952	0.887	0.89
C-HMCNN [37]	0.985	0.946	0.816	0.844	0.975	0.954	0.917	0.939	0.968	0.906	0.909
Chang et al. [4]	0.978	0.942	0.856	0.875	0.969	0.953	0.919	0.938	0.964	0.937	0.938
HRN [8]	0.987	0.955	0.866	0.886	0.975	0.958	0.926	0.945	0.974	0.936	0.938
HDL [9]	0.992	0.964	0.878	0.897	**0.979**	**0.961**	0.938	**0.952**	**0.977**	0.948	0.949
Ours (Resnet)	0.990	0.960	0.870	0.890	0.964	0.950	0.936	0.945	0.980	0.947	0.948
Ours (ViT)	**0.997**	**0.984**	**0.910**	**0.926**	0.970	0.956	**0.941**	0.950	0.968	**0.954**	**0.956**

Table 2: Performance comparison across three datasets.

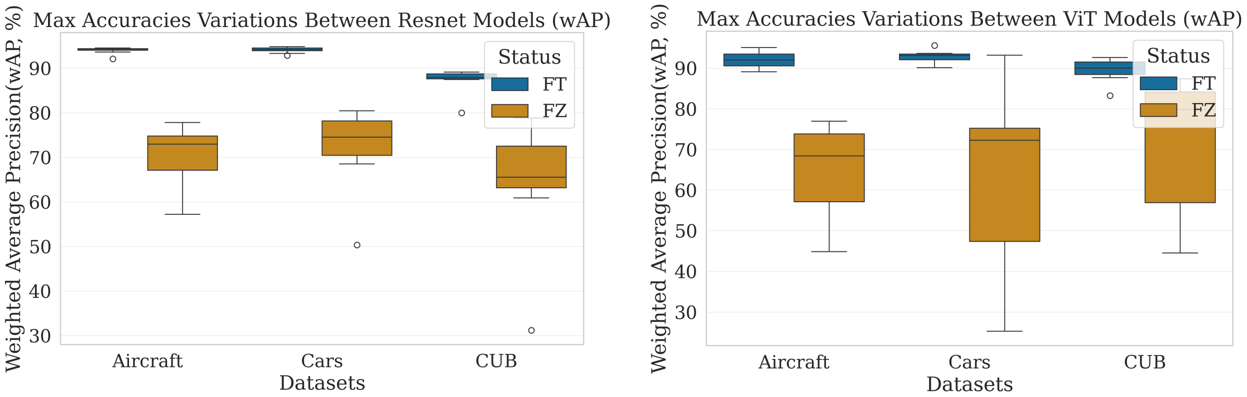

Fig. 3: Performance variability across backbone types and training modes. We compare fine-tuned (FT) and frozen (FZ) variants of ResNet (left) and ViT (right) models across hierarchical FGIR benchmarks.

els, this gap is worth noting as **frozen settings** only **train less than 0.05% of the trainable parameters**.

4.1.3. Backbone Choice Can Yield SOTA or Suboptimal Results

Figure 1 further illustrates the importance of backbone selection. For ResNets, top-performing models (e.g., MoCoV3) achieve near identical accuracy across datasets, Mocov3 obtains 94% wAP in Aircraft, while worst performing Resnet Spark 92% wAP. In contrast, ViT models show greater spread: SigLIP-v2 leads on Aircraft 94% wAP and Cars 95% wAP, while base ViT outperforms others on CUB 91% wAP. Worst ViT model is obtained by DINO with 87% wAP on Aircraft and 89% on Cars. These findings underscore that backbone selection, particularly for ViTs, can lead to either **state-of-the-art performance or significantly degraded results**.

4.2. Impact of Hyperparameters

4.2.1. Distribution of Results

In figure 4, Under **fine-tuned settings**, **ResNet models** pretrained via **discriminative models** (e.g., MoCoV3, SwAV) achieve the highest accuracy and lowest variance. Under **frozen setting**, **semi-supervised ResNet models** (e.g., IG1b, YFCC300m) and the IN21K-pretrained model outperform others, suggesting that pretraining dataset significantly influences performance under frozen settings.

For ViT models, Fine-Tuned Fully Supervised models such as base ViT, DeiT, and DeiT3 demonstrate both high mean accuracy and low standard deviation in fine-tuned settings. In the Frozen settings, Discriminative models like MoCoV3 and CLIP achieve the best performance.

4.2.2. Discriminative models perform best for both training settings

Discriminative models (e.g. MoCoV3, SwAV) obtains the best performance for both training settings, as observed in Fig. 4. Among the discriminative approaches, **contrastive methods outperform non-contrastive alternatives** (e.g. ViT DINO). This highlights the effectiveness of contrastive learning as a pretraining strategy for models deployed in structured, fine-grained classification pipelines. Discriminative models, trained via contrastive or predictive learning, is the best approach to tackle the core challenge of FGIR.

4.2.3. Size of pretraining Dataset influences frozen performance

Semi-Supervised models (IG1b, YFCC300m) and the **IN21K-P model achieve the best results** for ResNet **under the Frozen setting**. A common feature among these 3 models is that they are **pretrained on larger datasets, IG1b, YFCC300m, In21k accordingly**. For ViT models, where base ViT, DeiT (IN21K), and CLIP models achieve high accuracy, particularly on the CUB-200-2011 dataset.

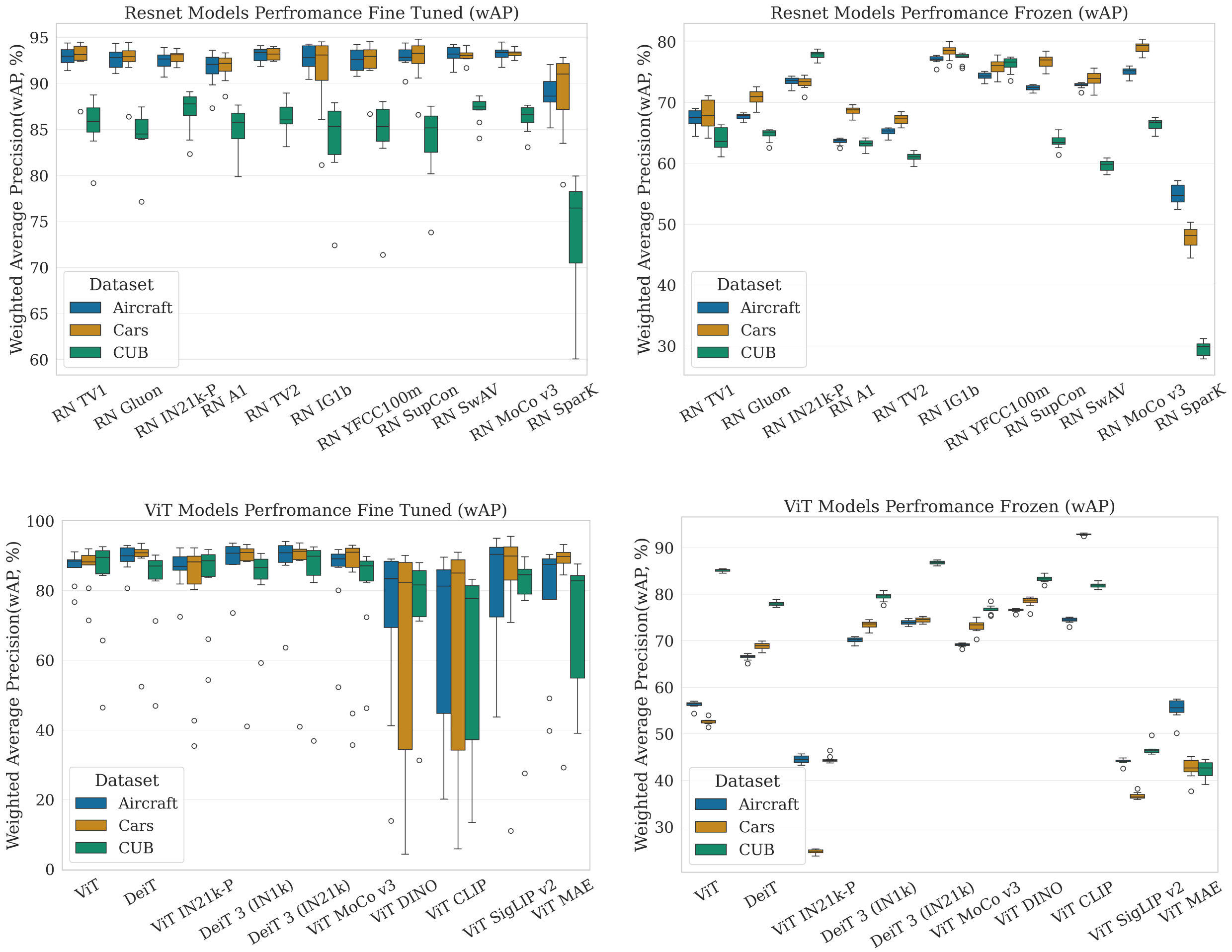

Fig. 4: Backbone performance across training regimes. We compare ResNet (top) and ViT (bottom) models under fine-tuned (FT, left) and frozen (FZ, right) settings. Each plot shows the wAP performance distribution across models with different pretraining strategies.

These models share the characteristic of pretrained on larger datasets, In21k, LAION5b accordingly, as observed in figure 4.

4.2.4. Generative models underperform in frozen settings

Generative models (e.g Resnet SparK, ViT MAE) consistently **underperform in frozen settings** for both ResNet and ViT backbones, as observed in Fig. 4. Generative models, trained via masking and reconstruction objectives, appear to struggle with the core challenge of distinguishing fine-grained categories, indicating that **this strategy is not well suited for frozen settings in hierarchical FGIR tasks**.

4.2.5. Resnet models do better on Aircraft, Cars, but ViT models do better on CUB

In figure (Fig. 1) and (Fig. 4)ResNet models tend to perform better on datasets like Aircraft and Cars, while ViT models perform better on CUB. This aligns with [5], which suggests that objects with more dynamic shapes, such as animals (represented in the CUB dataset), benefit more from the self-attention mechanism used in transformers, like ViT. In contrast, objects with more static shapes, such as aircraft and cars, benefit from the inductive biases inherent in CNN, like ResNet. Additionally, it is worth noting that ImageNet contains a substantial number of bird-related images, which may have contributed to a performance boost for ViT models on the CUB dataset. This supports the idea that the performance of these architectures vary depending on the type of objects in the dataset and the pretraining data they are exposed to.

4.3. Metric Correlation

Across all models, datasets, and evaluation targets, most transferability metrics—such as CKA, ImageNet-1k accuracy, and CIS [10, 11]—exhibit weak or inconsistent correlations with performance. As an illustrative example, Fig. 5 shows the correlation between a representative metric and Top-1 accuracy, highlighting the general trend: **no single metric reliably predicts downstream performance in the hierarchical FGIR setting**. While these metrics have demonstrated utility in standard transfer learning tasks, **they fail to generalize to hierarchical classification**, likely due to differences in label structure and the need to capture multilevel semantic relationships. This underscores a fundamen-

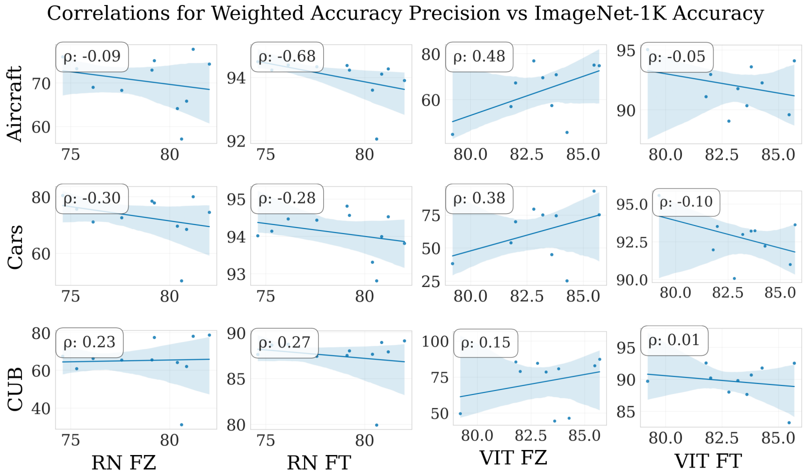

Fig. 5: Correlation between Imagenet-1k accuracy and wAP. Its correlation with downstream performance is weak, reflecting a broader trend observed across all tested metrics.

tal gap in existing metrics and motivates the development of more tailored transferability metrics for hierarchical FGIR.

5. CONCLUSION

The **choice of backbone** plays a **critical role** in hierarchical FGIR. By revisiting the pipeline of Chang et al. [4] with **systematic backbone and hyperparameter selection**, we achieve **state-of-the-art results**, even with **older architectures**. Our findings further underscore that **hyperparameter tuning exerts a substantial influence** on model performance, with **ResNet models more robust than ViTs** regardless of hyperparameters. In particular, **discriminative models** (e.g., MoCoV3, SwAV) yield strong performance with both **high accuracy and low variance**.

As for the predictive metrics, many proposed in prior works [10,11] and evaluate their generalizability to hierarchical FGIR. The majority of the metrics fail to consistently correlate with downstream performance across model families and training regimes. This challenges the existing assumptions in transferability research and underscores the need for metrics that account for hierarchical label structures and fine-grained task complexity.

REFERENCES

[1] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for Fine-grained Category Detection," July 2014. arXiv:1407.3867 [cs]. 1, 2

[2] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), pp. 5219–5227, IEEE, Oct. 2017. 1, 2

[3] T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, and L. Lin, "Fine-Grained Representation Learning and Recognition by Exploiting Hierarchical Semantic Embedding," Aug. 2018. arXiv:1808.04505 [cs]. 1, 2

[4] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your "Flamingo" is My "Bird": Fine-Grained, or Not," Mar. 2021. arXiv:2011.09040 [cs]. 1, 2, 3, 4, 5, 7

[5] S. Ye, Y. Wang, Q. Peng, X. You, and C. L. P. Chen, "The Image Data and Backbone in Weakly Supervised Fine-Grained Visual Categorization: A Revisit and Further Thinking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, pp. 2–16, Jan. 2024. 1, 6

[6] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple Granularity Descriptors for Fine-Grained Categorization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Santiago, Chile), pp. 2399–2406, IEEE, Dec. 2015. 1, 2

[7] R. Cerri, R. C. Barros, A. C. P. L. F. De Carvalho, and Y. Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC Bioinformatics*, vol. 17, p. 373, Sept. 2016. 1, 5

[8] J. Chen, P. Wang, J. Liu, and Y. Qian, "Label Relation Graphs Enhanced Hierarchical Residual Network for Hierarchical Multi-Granularity Classification," Jan. 2022. arXiv:2201.03194 [cs]. 1, 5

[9] H.-J. Chen, C.-J. Peng, H.-H. Shuai, and W.-H. Cheng, "A Hierarchically Discriminative Loss with Group Regularization for Fine-Grained Image Classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, p. 3698398, Oct. 2024. 1, 5

[10] S. Kornblith, J. Shlens, and Q. V. Le, "Do Better ImageNet Models Transfer Better?," June 2019. arXiv:1805.08974 [cs]. 2, 3, 6, 7

[11] N. Nayman, A. Golbert, A. Noy, T. Ping, and L. Zelnik-Manor, "Diverse Imagenet Models Transfer Better," Apr. 2022. arXiv:2204.09134 [cs]. 2, 3, 4, 6, 7

[12] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," June 2019. arXiv:1906.02243 [cs]. 2

[13] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning," Nov. 2019. arXiv:1910.09700 [cs]. 2

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 2021. arXiv:2010.11929 [cs]. 2

[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," Jan. 2021. arXiv:2012.12877 [cs]. 2

[16] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," in *Computer Vision – ECCV 2022* (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13684, pp. 516–533, Cham: Springer Nature Switzerland, 2022. Series Title: Lecture Notes in Computer Science. 2

[17] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K Pretraining for the Masses," Aug. 2021. arXiv:2104.10972 [cs]. 2, 3

[18] X. Chen, S. Xie, and K. He, "An Empirical Study of Training Self-Supervised Vision Transformers," Aug. 2021. arXiv:2104.02057 [cs]. 2, 3

[19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," May 2021. arXiv:2104.14294 [cs]. 2

[20] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, June 2023. arXiv:2212.07143 [cs]. 2

[21] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," Feb. 2025. arXiv:2502.14786 [cs]. 2

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," Dec. 2021. arXiv:2111.06377 [cs]. 2

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015. arXiv:1512.03385 [cs]. 2

[24] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," Dec. 2018. arXiv:1812.01187 [cs]. 2

[25] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," Oct. 2021. arXiv:2110.00476 [cs]. 3

[26] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," May 2019. arXiv:1905.00546 [cs]. 3

[27] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," Mar. 2021. arXiv:2004.11362 [cs]. 3

[28] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," Jan. 2021. arXiv:2006.09882 [cs]. 3

[29] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, and Z. Yuan, "Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling," Jan. 2023. arXiv:2301.03580 [cs]. 3

[30] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-Grained Visual Classification of Aircraft," June 2013. arXiv:1306.5151 [cs]. 3

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," 2011. 3

[32] J. Krause, J. Deng, M. Stark, and L. Fei-Fei, "Collecting a Large-Scale Dataset of Fine-Grained Cars," 3

[33] S. Kornblith, T. Chen, H. Lee, and M. Norouzi, "Why Do Better Loss Functions Lead to Less Transferable Features?," Nov. 2021. arXiv:2010.16402 [cs]. 3

[34] A. Islam, C.-F. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, "A Broad Study on the Transferability of Visual Representations with Contrastive Learning," Aug. 2021. arXiv:2103.13517 [cs]. 3

[35] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. 4

[36] J. Wehrmann, R. Cerri, and R. C. Barros, "Hierarchical Multi-Label Classification Networks," 5

[37] E. Giunchiglia and T. Lukasiewicz, "Coherent Hierarchical Multi-Label Classification Networks," Oct. 2020. arXiv:2010.10151 [cs]. 5