Fine Grained Image Classification Using Image Embedding Ensembles

Hüsamettin IŞIKTAŞ
Computer Engineering Department
Yıldız Technical University
İstanbul, Türkiye
husamettin.isiktas@std.yildiz.edu.tr

***Abstract*—Fine-Grained Visual Classification (FGVC) poses a significant challenge in computer vision due to high intra-class variance and low inter-class variance. While foundation models pre-trained on large-scale datasets have achieved remarkable success, single models often struggle to capture the full spectrum of discriminative features required for fine-grained tasks. This study investigates the effectiveness of "Image Embedding Ensembles" by combining diverse state-of-the-art architectures, including DINOv3, DINOv2, ConvNeXt V2, OpenCLIP, and SigLIP. We evaluate three ensemble strategies: Feature Concatenation (Early Fusion), Feature Summation, and Late Fusion (Logit Averaging), on the CUB-200-2011 and iNaturalist 2021 datasets. Our experiments demonstrate that ensemble methods significantly outperform individual models. Notably, the "Late Fusion" strategy combining all models achieved the best performance, boosting accuracy by approximately 7.6% on the challenging iNaturalist dataset compared to the best single model (DINOv3). These results highlight the collective power of diverse representations in solving complex classification problems.**

***Index Terms*—Fine-grained Visual Classification, Ensemble Learning, Transfer Learning, Image Embeddings, DINOv3, ConvNeXt, Late Fusion**

I. Introduction

Fine-Grained Visual Classification (FGVC) involves distinguishing between sub-ordinate categories such as bird species, car models, or plant varieties. Unlike generic object recognition, FGVC requires identifying subtle differences in local parts while being robust to significant variations in pose, background, and lighting. Recent advancements in deep learning have led to the development of powerful foundation models trained on massive datasets using self-supervised or weakly supervised learning. Models like DINOv2, OpenCLIP, and ConvNeXt have set new benchmarks in representation learning.

However, relying on a single model architecture may limit performance, as different models possess distinct inductive biases. For instance, Vision Transformers (ViTs) [1] excel at capturing global context through self-attention, while Convolutional Neural Networks (CNNs) are inherently strong at extracting local texture and shape information. This study proposes an ensemble approach that leverages these complementary strengths. We systematically evaluate the performance of five frozen backbones (DINOv3, DINOv2, ConvNeXt V2, OpenCLIP, SigLIP) and their combinations using distinct fusion strategies.

Our experiments on the CUB-200-2011 and iNaturalist 2021 datasets reveal that combining embeddings from diverse architectures yields superior accuracy and generalization compared to any single model. Specifically, we show that a "Late Fusion" of logits provides a robust and computationally efficient way to harness the collective intelligence of these foundation models.

II. Literature Review

Fine-Grained Visual Classification (FGVC) remains a pivotal challenge in computer vision, characterized by high intra-class variance and low inter-class variance [2]. Unlike generic object recognition, FGVC requires distinguishing between sub-categories such as bird species [3] or distinct biological taxa [4]. Recent advancements have shifted from specialized part-localization modules to leveraging large-scale Foundation Models (FMs) and ensemble strategies.

A. Convolutional Architectures and Modernization

Historically, Convolutional Neural Networks (CNNs) dominated FGVC through mechanisms like part-based modeling and attention-driven pooling to capture subtle discriminative features [5]. While Vision Transformers (ViTs) have gained prominence, CNNs have evolved to remain competitive. A notable development is **ConvNeXt V2** [6], which integrates a fully convolutional masked autoencoder (FCMAE) framework with Global Response Normalization (GRN). GRN mitigates feature collapse by enhancing inter-channel competition, a critical attribute for distinguishing fine-grained categories where feature diversity is essential. Empirical studies demonstrate that modern CNNs like ConvNeXt effectively capture local texture information often smoothed out by standard Transformers [6].

B. Self-Supervised Vision Transformers

The paradigm of Self-Supervised Learning (SSL) has revolutionized feature extraction by utilizing massive unlabeled datasets. The **DINO** family, specifically **DINOv3** [7], represents the state-of-the-art in this domain. Unlike its predecessors, DINOv3 scales to 7 billion parameters and employs "Gram Anchoring" regularization. This technique preserves high-quality dense feature maps during large-scale training, preventing the oversmoothing of patch tokens [7]. For FGVC tasks, this dense feature quality is paramount, as discriminative

cues are often localized (e.g., a beak or wing pattern). Benchmarks on iNaturalist 2021 and CUB-200 reveal that DINOv3's frozen features significantly outperform weakly supervised baselines [7], validating the utility of SSL for fine-grained tasks without extensive fine-tuning.

C. Vision-Language Alignment

Parallel to SSL, Vision-Language Models (VLMs) like CLIP [8] leverage natural language supervision to learn semantic representations. However, standard contrastive loss functions often struggle with the fine-grained distinctions required in FGVC due to noise in web-scale text data. **SigLIP** (Sigmoid Loss for Language Image Pre-training) [9] addresses this by replacing the softmax normalization with a pairwise sigmoid loss. This decoupling allows for larger effective batch sizes and more precise image-text alignment, resulting in superior zero-shot and few-shot performance on fine-grained benchmarks compared to standard CLIP models [9].

D. Ensemble and Hybrid Strategies

Given the complementary nature of different inductive biases, recent literature emphasizes ensemble learning. Hybrid architectures combining CNNs (for local features) and ViTs (for global context) have shown robustness in identifying weed species and medical anomalies [10]. Furthermore, "Late Fusion" strategies, which average the logits of diverse foundation models (e.g., combining DINO's structural understanding with SigLIP's semantic knowledge), have achieved state-of-the-art results in domains ranging from computational pathology [11] to biological taxonomy. These methods mitigate individual model biases, leveraging collective intelligence to resolve the variance paradox inherent in FGVC.

III. Methodology

In this study, to address the fine-grained visual classification problem, an "Image Embedding Ensembles" approach was adopted. The primary objective is to leverage the collective power of rich features (embeddings) obtained from different deep learning models pre-trained on large-scale datasets, rather than being limited by the representation capability of a single model. Accordingly, the aim is to maximize classification performance by combining features from models with diverse architectural structures using various ensemble strategies.

A. Datasets

Two fundamental datasets representing different difficulty levels of the problem were used in this study:

1) **CUB-200-2011 (Caltech-UCSD Birds-200-2011):** A standard benchmark dataset in the field of fine-grained classification. It contains a total of 11,788 images belonging to 200 different bird species. The dataset has a balanced distribution with approximately 60 images per class. In this study, the standard train/test split ('train_test_split.txt') provided with the dataset was considered.
2) **iNaturalist 2021 (Mini):** A large-scale dataset reflecting real-world conditions, containing high class imbalance and variation, captured in natural environments. Covering a total of 10,000 different species, the 'train_mini' subset was used for training, and the 'val' set was used for testing purposes. This dataset was selected to test the model's scalability and generalization capability. There was a 'test' dataset but it wasn't been labeled. So we divided train set as %80 train and %20 validation.

B. Feature Extraction and Models

To extract discriminative features from images, five State-of-the-Art (SOTA) models with different architectural principles and training objectives were utilized. A transfer learning approach was adopted for all models; their pre-training weights were frozen, and they were used solely as feature extractors (backbones).

- **DINOv3 (ViT-L/16):** Developed by Meta, this is one of the most up-to-date models in self-supervised learning literature. Based on the Vision Transformer (ViT-L/16) architecture, DINOv3 is trained with an approach combining masked image modeling and discriminative learning. The 'facebook/dinov3-vitl16-pretrain-lvd1689m' variation was used in this study.
- **DINOv2 (ViT-L/14):** The predecessor to DINOv3, this model has shown superior success in pixel-level tasks such as depth estimation and semantic segmentation [12]. The 'dinov2_vitl14' variation, producing a 1024-dimensional feature vector, was used. Thanks to its self-supervised nature, it learns robust visual representations without the need for labeled data.
- **ConvNeXt V2 (Large):** A modern architecture combining the strengths of Convolutional Neural Networks (CNNs) and Transformer models. Trained with the Masked Autoencoder (MAE) approach ('convnextv2_large.fcmae_ft_in22k_in1k'), this fully convolutional model effectively captures local texture and shape information.
- **OpenCLIP (ViT-L/14):** An open-source implementation of the CLIP (Contrastive Language-Image Pre-training) model, trained using contrastive learning with text-image pairs. Trained on the 'datacomp_xl_s13b_b90k' dataset, this model gains strong generalization capability by learning the relationship between visual content and semantic text information.
- **SigLIP (ViT-L/16):** A more efficient and scalable model using sigmoid loss (Sigmoid Loss for Language Image Pre-training) instead of the softmax loss found in standard CLIP training. The 'ViT-L-16-SigLIP-256' variation was used to evaluate the effect of image-text alignment on feature quality.

C. MLP Architecture

A specialized Multi-Layer Perceptron (MLP) network was designed to classify the extracted feature vectors. This network takes high-dimensional vectors from frozen feature extractors as input and produces class probabilities. The architectural details are as follows:

- **Input Layer:** Dynamically shaped according to the output dimension of the model (768-1536 for single usage, total dimension for combined usage).
- **Hidden Layers:** The model contains two hidden layers with 512 and 256 neurons, respectively. Each layer includes the ReLU activation function to learn nonlinear relationships, Batch Normalization [13] to stabilize training, and Dropout [14] at a rate of 0.5 to prevent overfitting.
- **Output Layer:** Contains neurons equal to the number of classes in the target dataset (200 for CUB, 10,000 for iNaturalist).

D. Ensemble Strategies

In this study, "Ensemble Learning" techniques were employed to minimize the errors of single models (weak learners) and increase generalization capability. The primary motivation is the assumption that models with different inductive biases can compensate for each other's weaknesses by responding differently to the same input. Specifically, the hybrid use of Transformer-based (using global attention mechanisms) and CNN-based (capturing local features) models maximizes this diversity. Three fundamental strategies were implemented:

*1) **Feature Concatenation (Early Fusion)**:* Vectors obtained from different feature extractors are concatenated end-to-end without any loss.

$$x_{concat} = [f_{DINOv3}(I), f_{DINOv2}(I), f_{ConvNeXt}(I)] \quad (1)$$

The dimension of the resulting feature vector is equal to the sum of the output dimensions of the models used ($d_{total} = \sum d_i$). The MLP classifier is trained on this enriched but high-dimensional space. The advantage of this method is that the model can access all features simultaneously; the disadvantage is the increased parameter count and the risk of the "curse of dimensionality" [15].

2) Feature Summation: Feature vectors from models are combined by element-wise summation.

$$x_{sum} = f_{DINOv3}(I) + f_{DINOv2}(I) + \dots \quad (2)$$

For this method to be applicable, all feature vectors must have the same dimension. Outputs of different dimensions (e.g., ConvNeXt) were first reduced to a common dimension (e.g., 1024) using a linear projection layer or PCA (Principal Component Analysis) [16]. The summation process provides a more compact representation by compressing the feature space (information compression) and reduces computational cost.

3) Late Fusion (Mean Ensembling): In this strategy, a separate MLP classifier is trained for each feature extractor. After training is complete, the outputs (logits) of the models are combined during the test phase. We employ an unweighted average of logits, preferred over class probabilities (Softmax) because logits contain the model's confidence level in a rawer form and are more resistant to outliers.

$$z_{final} = \frac{1}{N} \sum_{i=1}^{N} z_i \quad (3)$$

Here, z_i represents the logit output of the i-th model and N is the number of models. The final class prediction is made by an 'argmax' operation on the combined logit vector. This method allows for parallel training as it makes models independent of each other and prevents the failure of a single model from affecting the entire system.

IV. Experimental Results

For the classification task, a flexible Multi-Layer Perceptron (MLP) architecture was designed, independent of the feature extractors. This architecture operates adaptively regardless of the input dimension (single model or concatenated vectors):

- **Input Layer:** Accepts a d_{in} dimensional feature vector.
- **1st Hidden Layer:** A fully connected (Dense) layer with 512 neurons. It employs **ReLU** activation to learn nonlinear features, **Batch Normalization (1D)** for training stability, and **Dropout (p=0.5)** to prevent overfitting.
- **2nd Hidden Layer:** A fully connected layer with 256 neurons, configured similarly with ReLU, Batch Norm, and Dropout (p=0.5).
- **Output Layer:** A linear layer containing neurons equal to the number of classes (C) (200 for CUB, 10,000 for iNaturalist).

All models were trained on NVIDIA 5060Ti GPU hardware. The following hyperparameters were kept constant during the optimization process:

- **Optimization Algorithm:** AdamW [17]
- **Epochs:** 150 (No early stopping applied)
- **Batch Size:** 128
- **Learning Rate:** 0.001
- **Loss Function:** Cross-Entropy Loss (for multi-class classification)
- **Train/Val Split:** 20% of the training set was reserved for validation.

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score metrics. The model yielding the best validation F1 score during training was saved (Best Model Checkpointing).

A. Single Model Benchmarks

Before proceeding to ensemble experiments, the individual performances of five different feature extractors were compared on both datasets. This analysis played a role in determining the most suitable candidate models for ensemble strategies.

The results indicate that the **DINO** family (v3 and v2), trained with self-supervised learning, possesses superior representation capability on both CUB and iNaturalist datasets compared to other models. Particularly on a challenging dataset like iNaturalist, DINOv3 and DINOv2 outperformed their closest competitors, ConvNeXt and OpenCLIP, by approximately 7-8%. While ConvNeXt and OpenCLIP models

TABLE I
SINGLE MODEL PERFORMANCES ON CUB AND INATURALIST

Model	CUB Acc	CUB F1	iNat Acc	iNat F1
DINOv3	**0.897**	**0.897**	**0.722**	**0.719**
DINOv2	0.897	0.895	0.710	0.706
ConvNeXt	0.879	0.879	0.641	0.638
OpenCLIP	0.875	0.873	0.640	0.635
SigLIP	0.851	0.850	0.610	0.605

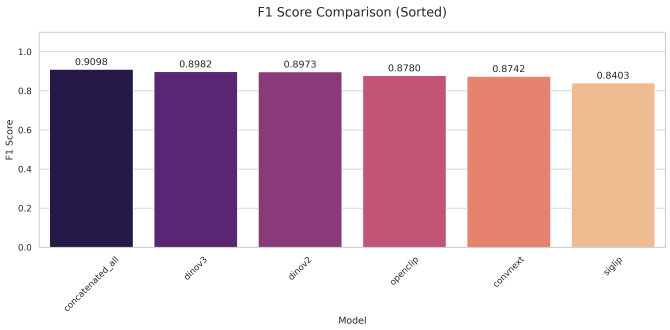

Fig. 1. F1-Score comparison of 5 different feature extractors on the CUB-200 dataset.

exhibited similar performances, SigLIP lagged behind in this task. In light of these findings, **DINOv3, DINOv2, and ConvNeXt** models, which showed the highest performance and provided architectural diversity, were selected for ensemble experiments.

B. CUB-200 Ensemble Learning Findings

Experiments conducted on the CUB dataset revealed that combining different architectures significantly increases success. Table II presents the results of all 10 experiments.

The results show that the "Late Fusion" method outperforms "Feature Concatenation" and "Feature Summation" methods.

C. iNaturalist 2021 Ensemble Learning Findings

In the iNaturalist dataset, where the number of classes and data imbalance are much higher, the effect of ensemble learning became more pronounced. Table III shows the results of all 9 experiments.

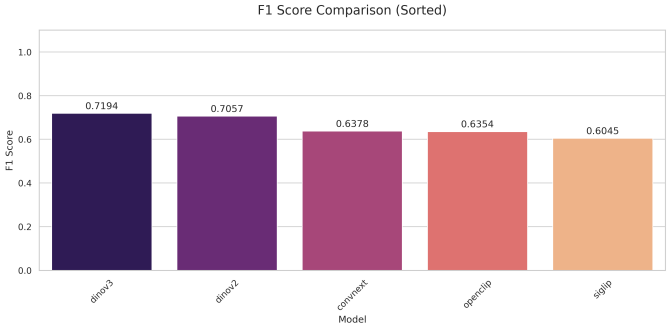

Fig. 2. F1-Score comparison of 5 different feature extractors on the iNaturalist 2021 dataset.

TABLE II
ENSEMBLE LEARNING PERFORMANCES ON CUB-200

Model / Method	Acc	F1-Score	Prec.	Recall
Ensemble All (Late Fusion)	**0.921**	**0.920**	**0.926**	**0.921**
Ens. (DINOv3 + DINOv2 + ConvNeXt)	0.918	0.917	0.922	0.918
Ens. (DINOv3 + DINOv2)	0.914	0.913	0.918	0.914
Ens. (DINOv3 + ConvNeXt)	0.911	0.911	0.917	0.911
Concat (DINOv3 + DINOv2)	0.910	0.908	0.920	0.910
Concat All	0.909	0.908	0.915	0.909
Concat (DINOv3 + ConvNeXt)	0.904	0.902	0.909	0.904
Sum (DINOv3 + DINOv2)	0.902	0.901	0.906	0.902
Single Reference (DINOv3)	*0.897*	*0.895*	*0.902*	*0.897*
Sum (DINOv3 + ConvNeXt)	0.892	0.892	0.898	0.892

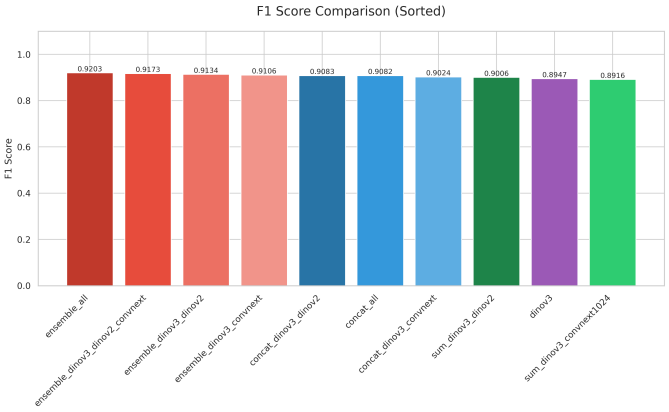

Fig. 3. F1-Score ranking of different ensemble learning strategies and single models on the CUB dataset.

On the iNaturalist dataset, while the single DINOv3 model remained at 72.2% accuracy, the "Ensemble All" method, which combines all models, achieved a success rate of 79.8%. This performance increase of approximately **7.6%** proves that as data complexity increases, single models become insufficient and collective intelligence plays a critical role. Training loss graphs indicate that the model converges stably despite the vast class space.

V. DISCUSSION AND CONCLUSION

In this study, we presented a comprehensive evaluation of "Image Embedding Ensembles" for Fine-Grained Visual Clas-

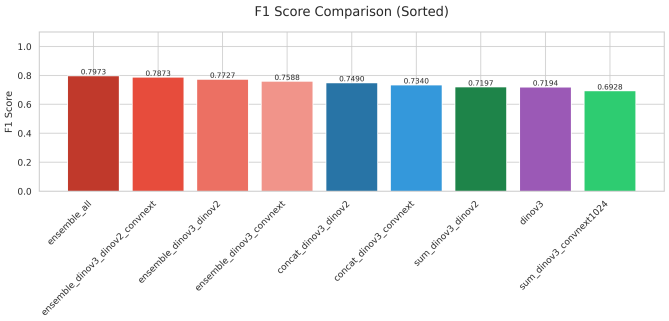

Fig. 4. F1-Score ranking of ensemble learning strategies and single models on the iNaturalist dataset.

TABLE III
ENSEMBLE LEARNING PERFORMANCES ON INATURALIST 2021

Model / Method	Acc	F1-Score	Prec.	Recall
Ensemble All (Late Fusion)	**0.798**	**0.797**	**0.816**	**0.798**
Ens. (DINOv3 + DINOv2 + ConvNeXt)	0.789	0.787	0.807	0.789
Ens. (DINOv3 + DINOv2)	0.775	0.773	0.794	0.775
Ens. (DINOv3 + ConvNeXt)	0.760	0.759	0.782	0.760
Concat (DINOv3 + DINOv2)	0.752	0.749	0.774	0.752
Concat (DINOv3 + ConvNeXt)	0.736	0.734	0.760	0.736
Sum (DINOv3 + DINOv2)	0.723	0.720	0.747	0.723
Single Reference (DINOv3)	*0.722*	*0.719*	*0.747*	*0.722*
Sum (DINOv3 + ConvNeXt)	0.696	0.693	0.722	0.696

sification, leveraging five state-of-the-art foundation models (DINOv3, DINOv2, ConvNeXt V2, OpenCLIP, and SigLIP). Our empirical findings offer several critical insights into the design of robust classification systems.

Firstly, the superiority of the "Late Fusion" strategy over high-dimensional feature concatenation (Early Fusion) is a significant finding. While concatenation provides the classifier with raw feature access, it suffers from the "curse of dimensionality" and increased optimization difficulty. Late Fusion, by operating on logits, effectively acts as a "committee of experts," where the independent confidence scores of diverse models smooth out individual prediction errors. This independence allows the ensemble to benefit from the unique inductive biases of each architecture without the interference of gradient conflicts during training.

Secondly, the performance disparity between the CUB-200 and iNaturalist datasets highlights the true value of ensemble learning. On CUB-200, single models already achieve near-saturation (approx. 90%), limiting the marginal gain of ensembles. However, on the iNaturalist dataset, which features a long-tail distribution and high environmental variance, the "Ensemble All" method yielded a substantial **7.6%** improvement over the best single model. This suggests that as the complexity and scale of the visual task increase, the "collective intelligence" of hybrid architectures (combining the global context of ViTs with the local texture bias of CNNs like ConvNeXt) becomes indispensable.

Despite these advantages, the computational overhead of running multiple large-scale backbones during inference remains a limitation for real-time applications. Future work will focus on two main directions: (1) **Knowledge Distillation**, where the high-performing ensemble serves as a teacher to train a lightweight student model, preserving accuracy while reducing latency; and (2) **Attention-based Fusion**, implementing learnable gates to dynamically weigh model contributions based on input difficulty, rather than simple averaging.

REFERENCES

[1] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[2] X. Wei *et al.*, "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, 2022.

[3] C. Wah *et al.*, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[4] G. Van Horn *et al.*, "The inaturalist species classification and detection dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8769–8778.

[5] J. Fu *et al.*, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4438–4446.

[6] S. Woo *et al.*, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16 133–16 142.

[7] O. Siméoni *et al.*, "Dinov3: Scaling self-supervised learning for vision foundation models," *arXiv preprint arXiv:2508.10104*, 2025.

[8] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[9] X. Zhai *et al.*, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11 975–11 986.

[10] L. Yuan *et al.*, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 558–567.

[11] M. Ochi and B. Yuan, "Ensemble of pathology foundation models for midog 2025 track 2: Atypical mitosis classification," *arXiv preprint arXiv:2509.02591*, 2025.

[12] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[14] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[16] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.