Few-Shot Classification of Fungi Species Using Contrastive Representation Learning and Multimodal Fusion

Lianping Lu[1], Heng Yang[1], Shuo Li[1], Fang Liu[1], Puhua Chen[1] and Wenping Ma[1]

[1]*Intelligent Perception and Image Understanding Lab, Xidian University*

Abstract

The FungiCLEF2025 challenge pioneers few-shot fungi species classification through multimodal observational data integration, specifically targeting the critical bottleneck of identifying rare and under documented taxa in practical biodiversity conservation scenarios. In this work, we present a novel two-stage framework that synergizes: (1) feature space optimization via Dynamic Weighting Contrastive Loss (DWCL), and (2) cross-modal fusion of visual characteristics with ecological metadata to achieve joint representation of environmental context and fine-grained morphological patterns. Through these technical innovations, the framework ultimately secured 2nd place in the competition leaderboard. The code is publicly available at https://github.com/Looploop555/fungi.

Keywords

FungiCLEF, Few-Shot Learning, Dynamic Weighting Contrastive Loss, Feature Fusion, Fine-grained Classification

1. Introduction

Fine grained visual categorization, a cornerstone challenge in computer vision and ecological informatics, holds critical implications for biodiversity monitoring and ecosystem conservation [1]. The FungiCLEF2025 challenge [2], co-hosted by CVPR-FGVC and LifeCLEF2025 [3], advances research on few-shot species recognition by leveraging multimodal observational data from real world citizen science initiatives. The FungiCLEF2025 challenge focuses on identifying fungi species under strictly limited training samples per class, where each taxon is defined by subtle morphological distinctions. Participants can integrate heterogeneous inputs, including multi view specimen imagery, geospatial coordinates, substrate, habitat annotations, and meteorological variables to discern fine-grained visual patterns critical for taxonomic differentiation. Building on previous FungiCLEF benchmarks [4, 5, 6], which demonstrated the efficacy of vision language models and metadata fusion techniques, this year's iteration introduces two core challenges:

- Few-shot feature learning: Learning discriminative representations from extremely limited training samples, with each class containing only 1-4 training instances.
- Multimodal data fusion: Jointly modeling specimen photographs with contextual metadata (e.g., spatiotemporal conditions, habitat descriptors) to amplify subtle taxonomic distinctions.

While previous FungiCLEF challenges have extensively explored open-set fungi classification paradigms, the critical challenge of few-shot classification remains notably underexamined. To address this issue, we investigate few-shot fungi recognition methods and report model performance under the competition's limited data conditions. Our framework incorporates two core innovations:

- Dynamic Weighting Contrastive Loss (DWCL): We introduce entropy-based uncertainty weighting and adaptive positive or negative pair construction, enabling robust intra class clustering and inter class separation even under few-shot conditions.
- Visual-Text Multimodal Fusion: Utilizing the Vision Transformer architecture [7], we conduct contrastive learning on DINOv2 [8] derived visual features to extract fine-grained visual patterns via its multi-head attention mechanism. Structured text generated from specimen metadata are

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

2323472564@qq.com (L. Lu); 1351519189@qq.com (H. Yang)

encoded using BERT and subsequently fused with visual features through Q-Former [9] based cross modal interaction.

- Two-Stage Decoupled Pipeline: By separating feature extraction and contrastive learning from multimodal fusion and final classification, each phase can be optimized independently. The first stage focuses on crafting highly discriminative visual embeddings, and the second stage integrates complementary modal signals.

2. Related Work

2.1. Fine-grained classification of Fungi

The participating teams in FungiCLEF2023 [5, 10, 11, 12], primarily employed Transformer-based [13] architectures for multimodal data processing, effectively combining visual features with metadata through advanced fusion strategies. To address critical challenges in fungi classification, the solutions incorporated specialized techniques including customized loss functions (such as Seesaw loss [14] and poisonous-classification loss) for handling class imbalance and long-tailed distributions.
The methods in FungiCLEF2024 [4, 15, 16, 17], primarily focused on multi-modal fusion of visual and metadata features using architectures like Swin Transformer V2 [18] and DINOv2, combined with dynamic MLPs [19] or attention mechanisms for fine-grained species classification. To handle open-set recognition, teams employed entropy-based rejection or generative adversarial approaches like OpenGAN [15] to detect unknown species. Safety-critical optimization was emphasized through poisonous-aware loss functions (e.g., heavily penalizing toxic misclassifications) and post hoc re-ranking to minimize dangerous errors. Auxiliary supervision (e.g., genus-level losses) and techniques like Seesaw Loss improved robustness against class imbalance.

2.2. Contrastive Learning

In the field of fine-grained classification, contrastive learning loss functions demonstrate unique advantages. Triplet Loss [20] constructs anchor-positive-negative triplets to enforce the distance between the anchor and the positive example to be smaller than that between the anchor and the negative example plus a margin. It aims to bring samples of the same class closer while pushing apart those of different classes, but its sampling efficiency is constrained by negative sample selection strategies. N-pair Loss [21] extends Triplet Loss by innovatively adopting a multi-negative parallel optimization mechanism, establishing a "1-positive-N-negative" contrast relationship within a single batch. However, when certain fungi categories have too few samples, their contribution as negative samples diminishes. Supervised Contrastive Loss [22] leverages supervised information to treat multiple samples from the same class as positives and those from different classes as negatives. It pulls same class samples closer in the embedding space while pushing apart different-class samples through contrastive learning. This approach is particularly suitable for supervised learning scenarios, excelling especially in few-shot learning and fine-grained classification tasks.

3. Method

We propose a two-stage framework for fine-grained fungi classification. In the first stage, foundational visual embeddings are extracted via DINOv2 and refined through a single layer Transformer encoder, then optimized with our Dynamic Weighting Contrastive, which incorporates entropy-based sample weighting and adaptive positive or negative pair construction to enhance intra class compactness and inter class separation even under scarce data regimes. In the second stage, we generate structured text from each specimen's metadata, encode them with BERT, and fuse the resulting text embeddings with the refined visual features using a Q-Former with a set of learnable queries q. This multimodal representation is trained with cross-entropy loss to produce habitat aware classification outputs, achieving competitive performance in FungiCLEF2025.

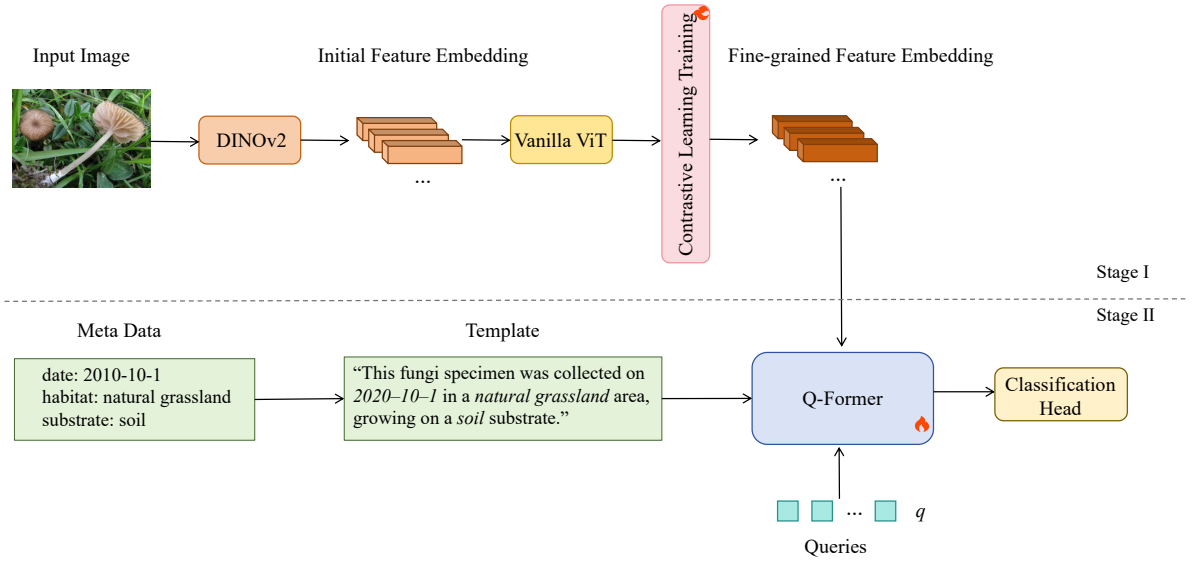

Figure 1: Overview of our method. Our framework employs a two-stage approach: first, DINOv2 extracts visual features optimized via Dynamic Weighting Contrastive Loss; then, BERT encoded metadata descriptions interact with visual features through Q-Former for cross modal fusion, ultimately producing classification predictions through a classifier.

3.1. Model Architecture

In the first stage, we extract initial visual features from each fungi image using DINOv2 and feed them into a Transformer based contrastive learning framework. This framework operates on pre-extracted features from a standard ViT and employs a single layer Transformer encoder with a 16 heads self attention mechanism to build a high dimensional attention space, effectively capturing fine-grained visual cues. In the second stage, for every fungi image, we construct a structured textual description template from its observation metadata-year, month, day, habitat, and substrate as follows:

> "This fungi specimen was collected on [year]–[month]–[day] in a [habitat] area, growing on a [substrate] substrate."

We design a two-stage model as shown in Figure 1. In the first stage, we concentrate on extracting and refining visual features; in the second stage, we carry out multimodal fusion and classification.

Subsequently, we employ BERT to encode the descriptions, then the generated text embeddings and the first stage visual features are jointly fed into the Q-Former module as input. Q-Former serves as the core component for cross modal fusion, establishing semantic relationships between image and ecological text descriptors. A set of learnable query tokens q. is introduced to facilitate cross modal interaction between textual and visual features. Through iterative updates via multi-head self attention, the Q-Former generates query representations that fuse habitat semantics with visual information. These representations are then projected through a classification head and optimized using cross-entropy loss to produce the final species classification results.

3.2. Training Strategy

In the first stage, we designed the Dynamic Weighting Contrastive Loss, an enhanced supervised contrastive loss function [22], which incorporates an entropy-based uncertainty weighting sampling mechanism to prioritize hard examples for optimized model training. Notably, our improvements to the standard loss function are as follows: First, uncertainty aware weighting: In loss calculation,

samples with higher prediction uncertainty are assigned greater weights, ensuring the model focuses on ambiguous instances critical for fine-grained discrimination. Second, adaptive pair construction: Positive pairs are formed by randomly sampling up to 4 instances per category, with a strict requirement of at least 2 samples per category to form valid pairs. For categories with fewer than 2 samples, new instances are generated via data augmentation to meet this constraint. Negative pairs are generated across distinct categories using a uniform class sampling strategy to avoid model bias. This design stabilizes the contrastive learning process by balancing positive and negative pairs while dynamically emphasizing samples that contribute most to reducing model uncertainty.

Given a batch of N samples, let $\mathbf{z}_i$ denote the feature vector of the i-th sample (including augmented instances for sparse categories). We first normalize the features:

$$\hat{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} \tag{1}$$

The pairwise similarity matrix is computed as:

$$S_{ij} = \hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_j^\top \tag{2}$$

The enhanced loss function is defined as:

$$\mathscr{L} = \frac{1}{\sum_{i\in\mathscr{A}} w_i} \sum_{i\in\mathscr{A}} w_i \cdot \frac{1}{|P(i)|} \sum_{j\in P(i)} -\log \frac{\exp(S_{ij}/\tau)}{\sum_{k\notin\mathscr{F}(i)} \exp(S_{ik}/\tau)} \tag{3}$$

where:

- τ is the temperature parameter.
- $P(i) = \{j \mid y_j = y_i, j \neq i\}$ denotes the set of positive samples for anchor i, with $|P(i)| \geq 2$ (augmented instances are included for sparse categories).
- $\mathscr{F}(i) = \{i\} \cup \{k \mid \text{mask}_k = 0\}$ represents invalid indices excluded by the triple masking mechanism (self-similarity and invalid pairs).
- $\mathscr{A} = \{i \mid |P(i)| \geq 2\}$ is the set of valid anchors.
- $w_i = \sigma(H(p_i))$ is the uncertainty weight for anchor i, where $H(p_i) = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c}$ (entropy of predicted probabilities p_i), and σ is the sigmoid function.

In the second stage, the text embeddings and visual features are integrated and fed into the Q-Former module. Meanwhile, the learnable query tokens are initialized as q. Q-Former performs interactive fusion between the textual and visual features through multi-head self attention, progressively updating the query tokens across multiple layers and representation subspaces to capture the fused multimodal information. The output query representations from Q-Former are then passed through a classification head for species prediction, and the final classification results are supervised using a cross-entropy loss function.

4. Experiment

4.1. Experimental Settings

Dataset. The FungiCLEF2025 challenge dataset is built from fungi observations submitted to the Atlas of Danish Fungi before the end of 2023, with labels provided by mycologists. It includes not only multiple photographs of the same specimen but also a wealth of supplementary data such as satellite imagery, meteorological records, and structured metadata. The vast majority of observations have been annotated with most of these attributes. As is shown in Table 1, The training set contains 4,293 observations, 7,819 images, and 2,427 classes, while the validation set has 1,099 observations, 2,285 images, and 570 classes. All of the images are also accompanied by tabular metadata and automatically-generated text descriptions of the images. Each class in the training set has between 1-4 observations.

Table 1
FungiCLEF2025 dataset statistics. The dataset exhibits limited sample sizes per category, with each class containing only a small number of images.

Subset	Observations	Species	All Images
Training	4293	2427	7819
Validation	1099	570	2285

Figure 2: Example of Fungi Images. The images capture the visual characteristics of fungi along with their ecological contexts.

As is shown in Figure 2, the images in this dataset primarily exhibit diverse visual characteristics of fungi and their growth environments, including mushroom close-ups, hyphal microstructures, and symbiotic surroundings. Most photographs employ tight close-up compositions that emphasize the spatial relationships between fungi and their substrates like decaying wood or soil. The striking color contrasts reflect both the complexity of natural field conditions and subtle biological morphological variations, providing visual data that combines macro-ecological context with micro-morphological features for fine-grained fungi classification.

Implementation Details. This method is developed based on the PyTorch framework [23]. The resolution of the input image is 224×224 pixels. We employ a 2048 dimensional embedding space, while effectively achieving feature disentanglement through a sophisticated 16 heads attention mechanism. All experiments are run on an H20-NVLink, using the AdamW optimizer [24]with a cosine annealing learning rate scheduler, and the initial learning rate set to 0.0002 and a batch size of 1024.

4.2. Evaluation Metric

The evaluation metrics for this competition is the standard Top@k which is defined as the proportion of instances where the true label is within the top k predicted labels:

$$\text{Top-}k \text{ Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{1}(y_i \in \hat{Y}_i^k)}{N}, \tag{4}$$

where:

- N is the total number of samples.
- y_i is the true label for the i-th sample.
- $\hat{Y}_i^k$ is the set of top k predicted labels for the i-th sample.
- $\mathbb{1}(\cdot)$ is the indicator function.

We set $k = 5$ for the main evaluation metric.

4.3. Fungi Dataset Experiments

As detailed in Table 2, when using only DINOv2 pretrained visual features, the model demonstrates relatively low Top5 accuracy, demonstrating that global visual features alone are insufficient for distinguishing morphologically similar fungi species. The incorporation of the Transformer encoder led to a significant improvement in accuracy, primarily attributed to the self-attention mechanism's dynamic focus on locally discriminative features. Further integration of habitat metadata boosted the model's accuracy to 76.991%, as the metadata provided complementary ecological information constraints to the visual features.

Table 2
When using only DINOv2 pretrained features, the Top5 accuracy reaches 37.128%, demonstrating that visual features alone are insufficient for capturing fine-grained distinctions among fungi specie. After incorporating a single layer Transformer encoder, the accuracy significantly improves to 70.892%, validating the effectiveness of self attention mechanisms in modeling local feature interactions. Further integration of habitat metadata yields an accuracy of 76.991%. The ecological prior knowledge provided by metadata synergizes with visual features, making the classification boundaries more distinct.

Method	Top5 Accuracy (%)
Only DINOv2	37.128
+Transformer	70.892
+Transformer+Metadata	**76.991**

Table 3
Impact of The Loss Function. Our proposed enhanced supervised contrastive loss DWCL achieves better performance. This is because the proposed DWCL enables the model to prioritize learning hard samples with ambiguous decision boundaries, thereby enhancing its capability to recognize challenging cases.

Loss Function	Top5 Accuracy (%)
Standard Contrastive Loss Function	74.778
Our(DWCL)	**76.991**

As detailed in Table 3, our enhanced loss function ensuring numerical robustness during training and delivering optimal performance in fine-grained fungi classification tasks. The Dynamic Weighted Contrastive Loss enhances the model's discriminative capability by focusing on challenging samples near decision boundaries, thereby improving classification performance for ambiguous cases.

Table 4
Ablation Study: Effects of ViT Depth, Number of Heads, and Training Epochs on Top-5 Accuracy. Our experiments reveal: (1) For the depth, a single layer architecture achieved optimal performance, as deeper models exhibited overfitting on the small-scale fungi dataset; (2) The 16 heads configuration outperformed 32 heads, demonstrating that moderate multi-head attention best captures fungi morphological subtleties; (3) The model achieves peak validation performance at epoch 100, while extending training to 150 epochs leads to a performance degradation, highlighting the necessity of early stopping strategies.

Experiment	Setting	Top5 Accuracy (%)
Depth	1	**76.991**
	2	72.566
Num Heads	16	**76.991**
	32	69.026
Epochs	50	75.221
	100	**76.991**
	150	75.663

As shown in Table 4, when training on small-scale datasets, excessively deep architectures may lead to overfitting, thereby reducing test set performance. The multi-head attention mechanism, as a core component of Transformer, captures richer feature information by simultaneously attending to different segments of the input sequence across multiple representation subspaces. In our experiments, the 16 heads configuration demonstrated superior performance compared to the 32 heads setup. The experimental results in Table 4 show that the model achieved high scores at 50, 100, and 150 training epochs. Building upon these three optimal results, we adopted a weighted voting ensemble approach [25] to integrate predictions from these top-performing models as our final competition submission. The aggregated final score reached 78.137%.

Table 5
Public leaderboard of FungiCLEF2025 competition(Partial). The proposed method ranks 2nd place.

Rank	Team	Top5 Accuracy (%)
1	Jack Etheredge	78.913
2	hard_work	78.137
3	aixiaodeyanjing	76.584
4	hahahahahal	76.196
5	skhhhh	75.291

The proposed two-stage framework, incorporating Dynamic Weighting Contrastive Loss for contrastive learning training and a multimodal data fusion strategy, achieved 2nd place on the official test set in the FungiCLEF2025 fine-grained few-shot fungi classification competition, as detailed in Table 5.

5. Conclusion

The proposed two-stage framework secured 2nd place in the FungiCLEF2025 competition. This achievement was accomplished through the integration of pretrained DINOv2 feature embeddings, a customized Transformer architecture, Dynamic Weighting Contrastive Loss, and metadata fusion strategies. Future research will focus on exploring satellite data augmentation and explainable attention mechanisms to facilitate practical field applications.

6. Declaration on Generative AI

During the preparation of this work, we did not use generative AI tools or services for writing assistance, figure generation, or data analysis. All text, figures, and results were produced solely by the authors.

References

[1] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 8927–8948.

[2] K. Janouskova, J. Matas, L. Picek, Overview of FungiCLEF 2025: Few-shot classification with rare fungi species, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[3] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2025.

[4] L. Picek, M. Šulc, J. Matas, Overview of fungiclef 2024: Revisiting fungi species recognition beyond 0-1 cost, in: CLEF 2024, 2024.

[5] L. Picek, M. Sulc, R. Chamidullin, J. Matas, Overview of fungiclef 2023: Fungi recognition beyond 1/0 cost., in: CLEF (Working Notes), 2023, pp. 1943–1953.
[6] L. Picek, M. Šulc, J. Heilmann-Clausen, J. Matas, Overview of FungiCLEF 2022: Fungi recognition as an open set classification problem, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
[9] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.
[10] H. Ren, H. Jiang, W. Luo, M. Meng, T. Zhang, Entropy-guided open-set fine-grained fungi recognition., in: CLEF (Working Notes), 2023, pp. 2122–2136.
[11] S. Wolf, J. Beyerer, Optimizing fine-grained fungi classification for diverse application-oriented open-set metrics., in: CLEF (Working Notes), 2023, pp. 2159–2167.
[12] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, Y. Li, X.-S. Wei, A deep learning based solution to fungiclef2023., in: CLEF (Working Notes), 2023, pp. 2051–2059.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[14] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, D. Lin, Seesaw loss for long-tailed instance segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9695–9704.
[15] J. Etheredge, Openwgan-gp for fine-grained open-set fungi classification, Working Notes of CLEF (2024).
[16] B.-F. Tan, Y.-Y. Li, P. Wang, L. Zhao, X.-S. Wei, Say no to the poisonous fungi: An effective strategy for reducing 0-1 cost in fungiclef2024, Training 1 (2024) 295–938.
[17] S. Wolf, P. Thelen, J. Beyerer, Poison-aware open-set fungi classification: Reducing the risk of poisonous confusion, Working Notes of CLEF (2024).
[18] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.
[19] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, J. Yang, Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10945–10954.
[20] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, IEEE (2015).
[21] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016, pp. 1857–1865. URL: https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.
[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673.
[23] A. Paszke, Pytorch: An imperative style, high-performance deep learning library, arXiv preprint arXiv:1912.01703 (2019).
[24] I. Loshchilov, F. Hutter, et al., Fixing weight decay regularization in adam, arXiv preprint arXiv:1711.05101 5 (2017) 5.
[25] L. Breiman, Bagging predictors, Machine learning 24 (1996) 123–140.