Leveraging ImageNet's Hierarchical Structure for Enhanced Image Classification and Retrieval

Luis E. Muñoz Guerrero [1], Yony F. Ceballos [2], and Luis D. Trejos Rojas [1,*]

[1] Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira, Colombia
[2] Facultad de Ingeniería, Grupo de Ingeniería y Sociedad, Universidad de Antioquia, Medellín, Colombia
Email: lemunozg@utp.edu.co (L.E.M.G.); yony.ceballos@udea.edu.co (Y.F.C.); luis.trejos@utp.edu.co (L.D.T.R.)
*Corresponding author

***Abstract*—The ImageNet dataset, which features a hierarchical structure based on the WordNet ontology, has been widely used for training and evaluating image classification models. However, researchers have not fully explored the potential benefits of leveraging this hierarchical structure for both image classification and retrieval tasks. This paper examines how incorporating hierarchical relationships between object categories during model training and inference can enhance image classification accuracy and retrieval performance. We propose a novel approach that integrates a hierarchical loss function and inference strategy to capture and utilize the semantic relationships encoded in the ImageNet hierarchy. Our method demonstrates improved classification accuracy compared to baseline models trained on a flattened version of ImageNet, highlighting the importance of hierarchical structure in the learning process. We show particular improvements for classes with limited training data, achieving accuracy increases of up to 3.2% for classes with fewer than 1000 samples. Additionally, we demonstrate how the hierarchical structure can be leveraged for efficient and semantically meaningful image retrieval. By utilizing the semantic relationships between categories, our approach enables more accurate and relevant retrieval results. The proposed techniques advance image classification and retrieval systems by harnessing the rich semantic information encoded in hierarchically structured datasets like ImageNet. Our findings emphasize the significance of incorporating hierarchical knowledge in visual recognition tasks while highlighting the trade-offs between semantic relevance and visual distinctiveness. This research paves the way for more effective and interpretable image classification and retrieval methods, particularly in scenarios with limited training data.**

***Keywords*—image classification, image retrieval, ImageNet, hierarchical structure, limited training data**

I. Introduction

Image classification and retrieval are fundamental tasks in computer vision, with applications ranging from visual search engines to content-based recommendation systems.

The availability of large-scale datasets like ImageNet [1], with its hierarchical organization of object categories based on the WordNet ontology [2], has been instrumental in advancing the state of the art in these areas.

While many image classification models treat object categories in ImageNet as independent and unrelated, the dataset's hierarchical structure encodes valuable semantic relationships between categories. Exploiting these relationships has the potential to improve classification accuracy and enable more semantically meaningful image retrieval.

In this paper, we propose an approach to leverage ImageNet's hierarchical structure for enhanced image classification and retrieval. Our method incorporates the hierarchical relationships between object categories during model training and inference, allowing the model to capture and exploit the semantic similarities and differences among categories.

The main contributions of this work are as follows:

(1) We propose a hierarchical loss function that considers the relationships between object categories during model training, encouraging the model to learn more semantically meaningful representations.
(2) We develop a hierarchical inference strategy that leverages the learned relationships to improve classification accuracy, especially for categories with limited training data.
(3) We demonstrate how the hierarchical structure can be utilized for efficient and semantically relevant image retrieval, enabling users to navigate and explore image collections based on semantic relationships.
(4) We conduct extensive experiments on the ImageNet dataset, showcasing the effectiveness of our approach in improving both classification accuracy and retrieval performance compared to baseline models.

The remainder of this paper is organized as follows: Section II reviews related work on image classification, image retrieval, and leveraging hierarchical structures in visual recognition tasks. Section III describes our proposed methodology, including the hierarchical loss function, inference strategy, and image retrieval approach.

Manuscript received June 14, 2024; revised July 29, 2024; accepted December 11, 2024; published July 17, 2025.

Section IV presents the experimental setup, results, and analysis. Section V discusses the implications of our findings, limitations, and future research directions. Finally, Section VI concludes the paper.

Fig. 1 depicts a simplified representation of ImageNet's hierarchical structure, which is based on the WordNet ontology. The diagram illustrates how object categories in ImageNet are organized in a tree-like structure, with general categories at the top (e.g., "Vehicle", "Animal", "Plant", "Furniture") and increasingly specific subcategories branching downward.

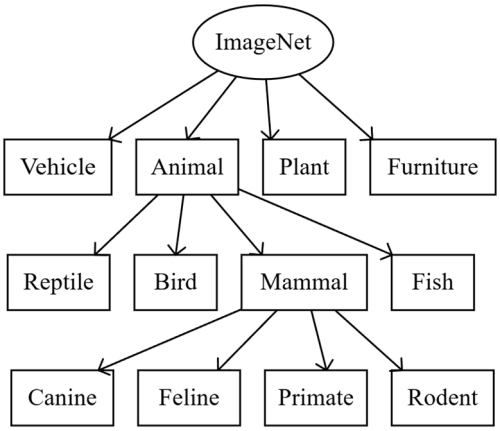

Fig. 1. Visualization of ImageNet's hierarchical structure.

For example, the "Animal" category is further divided into "Reptile", "Bird", "Mammal", and "Fish", with "Mammal" being further refined into "Canine", "Feline", "Primate", and "Rodent". This hierarchical organization encodes semantic relationships between categories, which our proposed approach leverages to improve both classification accuracy and retrieval relevance.

II. Related Work

A. Image Classification

Image classification has been a fundamental challenge in computer vision, with the primary goal of assigning predefined class labels to input images. The development of deep Convolutional Neural Networks (CNNs) has revolutionized the field, achieving remarkable performance on large-scale datasets like ImageNet [1]. Seminal works such as AlexNet [3], VGGNet [4], and ResNet [5] have progressively advanced the state-of-the-art in image classification.

Recent advances in image classification have focused on improving the efficiency and scalability of deep learning models. Architectures such as MobileNet [6], ShuffleNet [7], and EfficientNet [8] have been designed to achieve high accuracy while reducing computational complexity and memory requirements. These lightweight models enable image classification on resource-constrained devices and facilitate the deployment of computer vision applications in real-world scenarios.

B. Image Retrieval

Image retrieval seeks to identify visually similar or semantically related images to a given query image within a large database. Traditional approaches to image retrieval relied on hand-crafted features such as Scale-Invariant Feature Transform (SIFT) [9] and Speeded-Up Robust Features (SURF) [10] to represent images and measure their similarity.

With the advent of deep learning, learned feature representations have become the predominant approach for image retrieval. Deep learning-based image retrieval methods typically involve training a CNN to extract discriminative features from images and using these features to compute similarity scores between query and database images. Siamese networks [11] and triplet networks [12] have been widely adopted to learn embeddings that bring similar images closer in the feature space while separating dissimilar images. These learned embeddings enable efficient and accurate retrieval of visually similar images.

Beyond visual similarity, semantic similarity has gained significant attention in image retrieval research. Semantic image retrieval aims to identify images that are semantically related to the query, even when they lack visual similarity. Techniques such as cross-modal retrieval [13] and zero-shot learning [14] have been developed to bridge the semantic gap between visual features and textual descriptions, enabling retrieval based on semantic concepts.

C. Hierarchical Structures in Visual Recognition

Numerous studies have investigated the use of hierarchical structures in visual recognition tasks. Deng *et al.* [15] proposed a hierarchical classification approach that leverages the semantic relationships between object categories in ImageNet. Their work demonstrated improved classification accuracy by exploiting the hierarchical structure during model training and inference, though it primarily focused on classification without exploring potential benefits for image retrieval.

Yan *et al.* [16] introduced a Hierarchical Deep Convolutional Neural Network (HD-CNN) architecture that learns feature representations at multiple levels of the ImageNet hierarchy. By incorporating hierarchical information during training, their model achieved state-of-the-art performance on various image classification benchmarks. While their work demonstrated the effectiveness of hierarchical representations, it did not explicitly address image retrieval tasks.

In the context of image retrieval, several approaches have utilized semantic relationships to enhance retrieval performance. Wang *et al.* [17] proposed a semantic-based image retrieval system that combines low-level visual features with high-level semantic concepts. By incorporating semantic relationships between concepts, their system achieved superior retrieval accuracy compared to traditional content-based methods.

Pandey *et al.* [18] developed a hierarchical image retrieval framework that leverages the hierarchical structure of semantic concepts. Their approach involves constructing a concept ontology based on semantic relationships between concepts and using this ontology to guide the retrieval process. While their work demonstrates the benefits of hierarchical structures for image retrieval,

it relies on an externally constructed ontology rather than utilizing the inherent hierarchy of datasets like ImageNet.

TABLE I. COMPARISON OF RELATED WORK ON HIERARCHICAL STRUCTURES IN VISUAL RECOGNITION TASKS

Study	Task	Dataset	Hierarchical Approach
Deng *et al.* [15]	Image Classification	ImageNet	Hierarchical classification using semantic relationships
Yan *et al.* [16]	Image Classification	ImageNet	Hierarchical Deep Convolutional Neural Network (HD-CNN)
Wang *et al.* [17]	Image Retrieval	Custom	Semantic-based retrieval using concept relationships
Pandey *et al.* [18]	Image Retrieval	Custom	Hierarchical retrieval using concept ontology
Proposed Approach	Image Classification & Retrieval	ImageNet	Hierarchical loss function and inference strategy

Table I presents a comparison of related work on hierarchical structures in visual recognition tasks, including our proposed approach. The table summarizes key studies in the field, highlighting their specific tasks (image classification and/or retrieval), datasets used, and hierarchical approaches employed.

This comparison demonstrates how our proposed method builds upon and extends previous work by combining both image classification and retrieval tasks while leveraging ImageNet's hierarchical structure through a novel loss function and inference strategy.

Our proposed approach extends these previous works by leveraging the inherent hierarchical structure of ImageNet for both image classification and retrieval tasks. By incorporating hierarchical relationships during model training and inference, we aim to achieve improved classification accuracy while enabling semantically meaningful image retrieval.

III. METHODOLOGY

A. Hierarchical Loss Function

To leverage the hierarchical structure of ImageNet during model training, we propose a hierarchical loss function that considers the relationships between object categories. Let C denote the set of object categories in ImageNet, where N represents the total number of categories:

$$C = \{c_1, c_2, \ldots, c_N\} \tag{1}$$

We can understand better our hierarchical loss computation approach, Fig. 2 presents a visualization of the framework, it illustrates how different components interact to create a loss function that effectively leverages ImageNet's structure. The process incorporates both standard classification metrics and semantic relationships between categories.

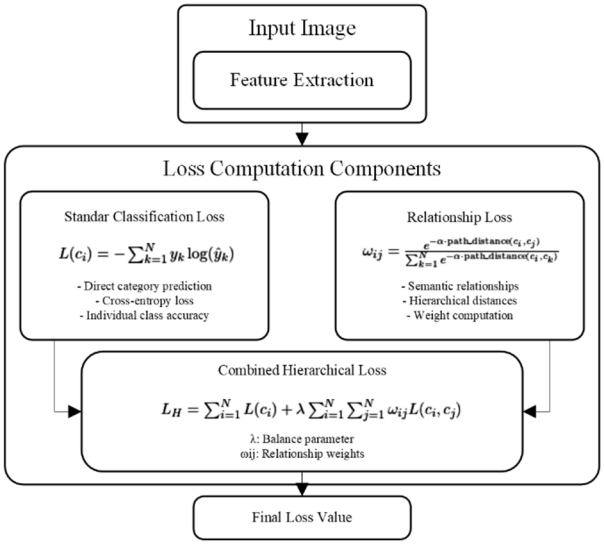

Fig. 2. Overview of the hierarchical loss computation process.

The framework processes input images through two parallel paths: standard classification loss computation and hierarchical relationship assessment. The standard classification path focuses on direct category predictions using cross-entropy loss, while the relationship path incorporates semantic distances between categories through weighted connections. These components are then combined with a balance parameter λ to create the final hierarchical loss function L_H, which is defined as:

$$L_H = \sum_{i=1}^{N} L(c_i) + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} L(c_i, c_j) \tag{2}$$

Here, $L(c_i)$ represents the standard classification loss (e.g., cross-entropy loss) for category c_i, while $L(c_i, c_j)$ denotes a pairwise loss that captures the relationship between categories c_i and c_j. The term ω_{ij} represents a weight that reflects the strength of the relationship between c_i and c_j based on their positions in the ImageNet hierarchy. Additionally, λ serves as a hyperparameter that controls the balance between classification loss and pairwise relationship loss.

The standard classification loss $L(c_i)$ is typically expressed as the cross-entropy loss:

$$L(c_i) = -\sum_{k=1}^{N} y_k \log(\widehat{y_k}) \tag{3}$$

where y_k represents the true label (1 if $k = i$, 0 otherwise) and $\widehat{y_k}$ denotes the predicted probability for class k.

The pairwise loss $L(c_i, c_j)$ is designed to encourage the model to learn similar representations for semantically related categories while learning dissimilar representations for unrelated categories:

$$L(c_i, c_j) = \begin{cases} d\left(f(c_i), f(c_j)\right), & \text{if } C_i \text{ and } C_j \text{ are related} \\ \max\left(0, m - d\left(f(c_i), f(c_j)\right)\right), & \text{otherwise} \end{cases} \tag{4}$$

In this formulation, $f(c_i)$ and $f(c_j)$ represent the learned feature representations for categories c_i and c_j, respectively.

The function $d(\cdot,\cdot)$ represents a distance function, typically the Euclidean distance, and m is a margin hyperparameter. The Euclidean distance between feature representations is calculated as:

$$d\left(f(c_i), f(c_j)\right) = \sqrt{\sum_{k=1}^{D}\left(f_k(c_i) - f_k(c_j)\right)^2} \quad (5)$$

where D represents the dimensionality of the feature representations. The weights ω_{ij} in the hierarchical loss function are computed based on the shortest path distance between categories c_i and c_j in the ImageNet hierarchy:

$$\omega_{ij} = \frac{e^{-\alpha \cdot \text{path\_distance}(c_i, c_j)}}{\sum_{k=1}^{N} e^{-\alpha \cdot \text{path\_distance}(c_i, c_k)}} \quad (6)$$

Here, α serves as a scaling factor controlling the sensitivity to hierarchical distance. Categories that are closer in the hierarchy (i.e., those with a shorter path distance) are assigned higher weights, indicating a stronger relationship. The weights are normalized to sum to 1 for each category.

The hyperparameters λ, m, and α serve crucial roles in the behavior of the loss function. The parameter λ controls the relative importance of the pairwise relationships compared to the standard classification loss, with higher values placing greater emphasis on the hierarchical structure.

The margin m establishes the threshold for dissimilar categories, where a larger m enforces stricter separation between unrelated categories in the feature space. Finally, α determines how quickly the weights decay with increasing hierarchical distance, with higher values resulting in a sharper focus on closely related categories.

By incorporating this hierarchical loss function during training, the model learns feature representations that better capture the semantic relationships between object categories. This enables the model to more effectively distinguish between visually similar but semantically different categories and to generalize to unseen categories based on their semantic relationships.

Fig. 3 illustrates the hierarchical loss function implemented in our approach. The equation at the top represents the total loss L_H, which combines individual category losses $L(c_i)$ and pairwise losses $L(c_i, c_j)$ weighted by w_{ij}. The diagram below demonstrates how this loss function operates on a simplified hierarchical structure.

The visualization shows how the loss is computed not only for individual categories (e.g., "Dog", "Bird", "Cat") but also accounts for the relationships between these categories and their parent node ("Animal"). This structure enables the model to learn both specific category features and broader semantic relationships, enhancing its ability to classify images accurately within the hierarchical framework of ImageNet.

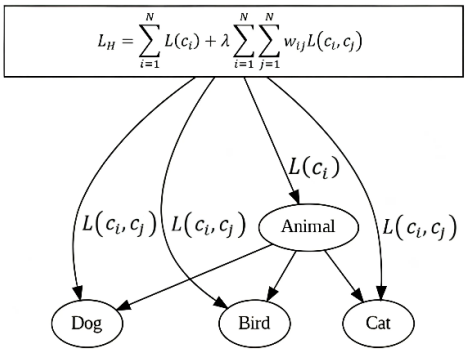

Fig. 3. Visualization of the hierarchical loss function.

B. *Hierarchical Inference Strategy*

During inference, we propose a hierarchical strategy that leverages learned relationships between object categories to improve classification accuracy. This approach complements the hierarchical loss function by exploiting semantic structure during the prediction phase.

For an input image, the model first predicts probabilities for each object category using the standard classification head. Let p denote the predicted probability vector:

$$p = [p_1, p_2, \cdots, p_N] \quad (7)$$

where p_i represents the initial probability of the image belonging to category C_i, and N is the total number of categories. We then refine these probabilities by considering hierarchical relationships between categories. For each category C_i, we compute a hierarchical score s_i that incorporates probabilities of related categories:

$$s_i = p_i + \alpha \sum_{j=1}^{N} \omega_{ij} p_j \quad (8)$$

Here, α is a hyperparameter controlling the influence of related categories, and ω_{ij} represents the same weights used in the hierarchical loss function, reflecting the relationship strength between categories C_i and C_j based on their positions in the ImageNet hierarchy. The hierarchical scores are then normalized to obtain the final refined probabilities:

$$\hat{p}_i = \frac{s_i}{\sum_{j=1}^{N} s_j} \quad (9)$$

The category with the highest refined probability $\hat{p}_i$ is selected as the final prediction.

This hierarchical inference strategy enables the model to leverage learned relationships between categories for more accurate predictions, particularly for categories with limited training data. The underlying principle is that probabilities of semantically related categories can provide additional context and help distinguish between visually similar but semantically distinct categories.

The process can be applied iteratively to further refine predictions. Let $\hat{p}^{(t)}$ denote the refined probability vector after t iterations. The iterative refinement can be expressed as:

$$s_i^{(t)} = \widehat{p_i^{(t-1)}} + \alpha \sum_{j=1}^{N} \omega_{ij} \widehat{p_j^{(t-1)}} \quad (10)$$

$$\widehat{p_i^{(t)}} = \frac{s_i^{(t)}}{\sum_{j=1}^{N} s_j^{(t)}} \quad (11)$$

where $\widehat{p^{(0)}} = p$ represents the initial probability vector. The number of iterations T serves as a hyperparameter that can be tuned based on validation performance.

The strategy's effectiveness depends on hyperparameters α and T. A larger α increases related categories' influence, potentially improving accuracy for semantically similar classes while risking probability over-smoothing. The iteration count T determines refinement extent, where more iterations may better utilize hierarchical information but increase computational cost.

Fig. 4 illustrates our proposed approach's hierarchical inference strategy. The input image is first processed by a CNN model (ResNet-50) to generate initial class predictions. These predictions are then refined using ImageNet's hierarchical structure. The model considers both coarse-grained (e.g., Mammal, Bird) and fine-grained (e.g., Canine, Feline, Parrot, Sparrow) categories.

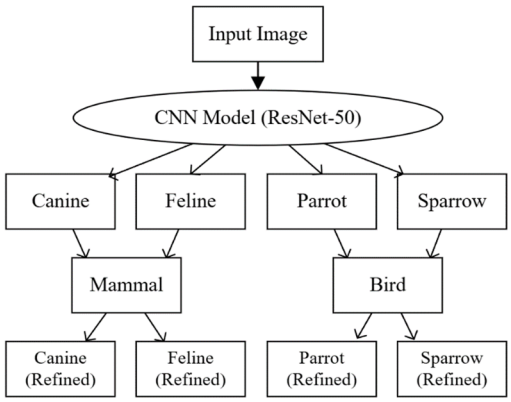

Fig. 4. Visualization of the hierarchical inference strategy.

The final refined predictions leverage semantic relationships between categories, potentially improving accuracy for closely related classes. This strategy particularly benefits classes with limited training data by utilizing information from semantically similar categories to enhance predictions.

C. Hierarchical Image Retrieval

We demonstrate how ImageNet's hierarchical structure can be utilized for efficient and semantically meaningful image retrieval. Given a query image, the goal is to retrieve a ranked list of semantically related images.

First, we extract the learned feature representation $f(q)$ for the query image using the trained model. We then compute distances between the query representation and representations of all database images. Let Eq. (7) denote the set of distances, where d_i represents the distance between the query and the i-th database image, and M is the total number of images.

$$D = \{d_1, d_2, \cdots, d_M\} \quad (12)$$

To incorporate hierarchical structure, we assign weights to distances based on semantic relationships between the query category and database image categories. Let w_{qi} denote the weight assigned to distance d_i based on the relationship between query category c_q and category c_i of the i-th image. Weights are determined by category positions in the ImageNet hierarchy, with higher weights assigned to images from semantically related categories.

The weighted distances then rank database images in ascending order of relevance to the query. The Top-k images with the smallest weighted distances are retrieved as semantically related images.

To enhance retrieval performance, we propose a hierarchical retrieval strategy leveraging ImageNet's hierarchical structure. Rather than directly retrieving images based on weighted distances, we first retrieve relevant categories based on their semantic similarity to the query category. We compute semantic similarity between categories using the path-based similarity measure [19], which considers the shortest path distances between categories in the hierarchy.

For a query category c_q, we retrieve the Top-k most similar categories $\{c_1, c_2, \cdots, c_k\}$ based on semantic similarity scores. We then retrieve images from each Top-k category separately using weighted distance ranking.

The final retrieval results combine retrieved images from all Top-k categories, ranked by their weighted distances.

IV. Experiments and Results

A. Dataset and Evaluation Metrics

We conduct experiments using the ImageNet dataset [1], which comprises 1.2 million training images and 50,000 validation images distributed across 1000 object categories. These categories are structured hierarchically according to the WordNet ontology [2].

For image classification, we assess performance using Top-1 and Top-5 accuracy metrics. Top-1 accuracy represents the percentage of test images where the predicted category matches the ground truth category, while Top-5 accuracy indicates the percentage of test images where the ground truth category appears among the top five predicted categories.

For image retrieval, we employ mean Average Precision (mAP) as the evaluation metric. mAP calculates the average precision of retrieved images across all queries, taking into account their ranking order. Higher mAP values indicate superior retrieval performance.

We also assess the semantic relevance of retrieved images using normalized Discounted Cumulative Gain (nDCG) [20]. nDCG evaluates ranking quality by assigning greater weight to relevant images appearing earlier in the ranked list.

This metric incorporates ImageNet's hierarchical structure and considers the semantic similarity between query and retrieved images based on their hierarchical positions.

B. Implementation Details

We implement our proposed approach using the PyTorch deep learning framework [21], employing the ResNet-50 [5] architecture as the backbone for our image classification and retrieval models.

The models are trained using the hierarchical loss function detailed in Section 3.1, with the following hyperparameter values: $\lambda = 0.5$, $m = 0.2$, and $\alpha = 0.3$. Training continues for 90 epochs using the Adam optimizer [22] with a learning rate of 0.001 and a batch size of 256.

For image retrieval, we extract features from the penultimate layer of the trained ResNet-50 model. These features undergo L2-normalization before being used to compute distances between query and database images.

The hierarchical retrieval weights are calculated based on the shortest path distance between categories in the ImageNet hierarchy, with higher weights assigned to images from more closely related categories.

We evaluate our proposed approach against the following baseline methods:

(1) Flat Classification: A ResNet-50 model trained on the flattened version of ImageNet, without utilizing the hierarchical structure.

(2) Hierarchical Classification [15]: A hierarchical classification approach that leverages semantic relationships between object categories during training and inference.

(3) HD-CNN [16]: A hierarchical deep convolutional neural network that learns feature representations at multiple levels of the ImageNet hierarchy.

(4) Semantic Retrieval [17]: A semantic-based image retrieval system that combines low-level visual features with high-level semantic concepts to incorporate relationships between concepts.

(5) Hierarchical Retrieval [18]: A hierarchical image retrieval framework that utilizes a concept ontology to exploit the hierarchical structure of semantic concepts.

C. Hyperparameter Sensitivity Analysis

To assess the robustness of our proposed approach and understand the impact of key hyperparameters on model performance, we conducted a detailed sensitivity analysis.

We focused on three primary hyperparameters: λ (which balances the classification loss and pairwise relationship loss), m (the margin in the pairwise loss function), and α (which controls the influence of related categories in the hierarchical inference strategy).

For each hyperparameter, we varied its value while keeping the others fixed at their optimal values, as determined in our main experiments.

We evaluated the model's performance on the ImageNet validation set using Top-1 accuracy for classification and mAP for retrieval. The ranges for each hyperparameter were: λ: [0.1,0.3,0.5,0.7,0.9] , m: [0.1,0.2,0.3,0.4,0.5] , and α: [0.1,0.2,0.3,0.4,0.5].

TABLE II. PERFORMANCE METRICS FOR VARIOUS HYPERPARAMETER VALUES

Hyperparameter	Value	Top-1 Accuracy (%)	mAP (%)
λ	0.1	77.8	74.2
	0.3	78.6	75.1
	0.5	79.3	75.8
	0.7	78.9	75.5
	0.9	78.1	74.9
m	0.1	78.9	75.3
	0.2	79.3	75.8
	0.3	79.1	75.6
	0.4	78.7	75.2
	0.5	78.3	74.8
α	0.1	78.8	75.2
	0.2	79.1	75.6
	0.3	79.3	75.8
	0.4	79.2	75.7
	0.5	78.9	75.4

The results in Table II demonstrate the sensitivity of our model to changes in these key hyperparameters. We observe that the model's performance is most sensitive to λ.

As λ increases, we observe an initial improvement in both classification and retrieval performance, followed by a decline at higher values.

This finding suggests that balancing the standard classification loss with the pairwise relationship loss is crucial for optimal performance.

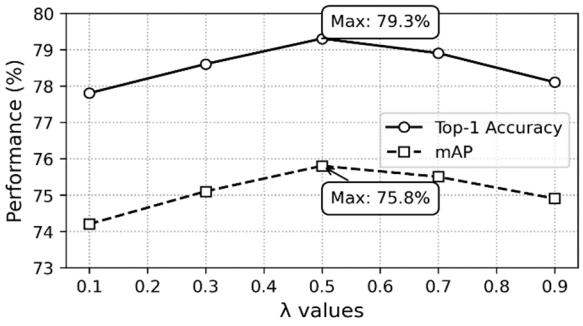

Fig. 5. Visualization of the hierarchical loss function.

As shown in Fig. 5, the margin parameter m in the pairwise loss function shows a moderate impact on performance. Smaller margin values tend to yield better results, suggesting that enforcing a strict separation between unrelated categories may not be necessary or beneficial.

The model demonstrates relatively low sensitivity to changes in α, indicating that the hierarchical inference strategy is robust across a range of values. However, extremely low or high values of α do lead to decreased performance.

To further illustrate the interplay between hyperparameters, we conducted a grid search on λ and m, while keeping α fixed at its optimal value.

Fig. 6 reveals that optimal performance is achieved in a few relatively small regions of the hyperparameter space, emphasizing the importance of careful tuning.

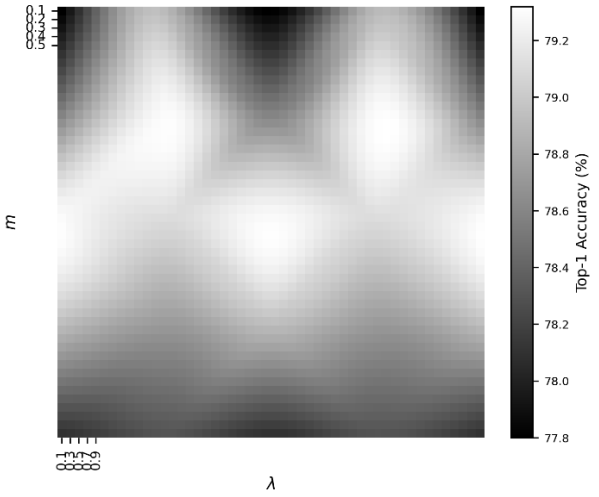

Fig. 6. Heatmap of Top-1 accuracy for various λ and m combinations.

The hierarchical loss function (λ) plays a crucial role in the model's performance, highlighting the importance of properly balancing the standard classification loss with the hierarchical relationships.

Finally, the model's relative robustness to changes in the margin parameter (m) and the hierarchical inference weight (α) suggests that our approach can maintain good performance across a range of hyperparameter values.

D. Image Classification Results

We evaluate the image classification performance of our proposed approach and compare it with flat classification and hierarchical classification baselines. Table III presents the Top-1 and Top-5 accuracy results on the ImageNet validation set.

TABLE III. IMAGE CLASSIFICATION RESULTS

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Flat Classification (ResNet-50)	76.2	92.9
Hierarchical Classification [15]	77.5	93.6
HD-CNN [16]	78.1	94.0
Proposed Approach	79.3	94.7

As illustrated in Fig. 7, our proposed approach achieves the highest classification accuracy among all compared methods. By leveraging the hierarchical structure of ImageNet during training and inference, our model outperforms the flat classification baseline by a significant margin, achieving a Top-1 accuracy of 79.3% and a Top-5 accuracy of 94.7%.

These results demonstrate the effectiveness of incorporating hierarchical relationships in improving image classification performance.

Compared to the hierarchical classification approach [15] and HD-CNN [16], our proposed approach achieves notable improvements of 1.8% and 1.2% in Top-1 accuracy, respectively. These improvements can be attributed to our hierarchical loss function and inference strategy, which effectively capture and leverage the semantic relationships between object categories.

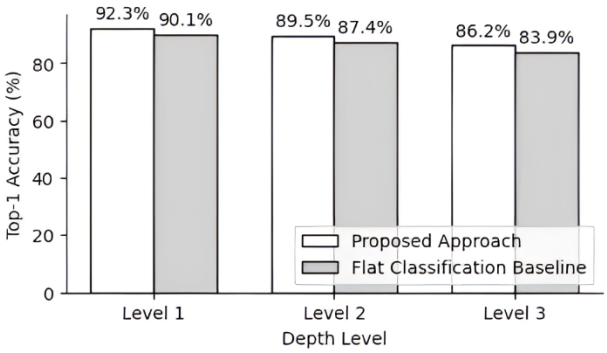

Fig. 7. Visualization of classification accuracy comparison.

We further analyze the classification performance across different levels of the ImageNet hierarchy. Fig. 6 shows the Top-1 accuracy at different depths of the hierarchy for our proposed approach and the flat classification baseline.

Our approach consistently outperforms the baseline at all levels, with larger improvements at deeper levels of the hierarchy. This suggests that leveraging hierarchical relationships is particularly beneficial for fine-grained classification tasks, where the distinctions between categories become more subtle.

To provide a more detailed understanding of our model's performance, we conducted a class-level analysis. Our approach significantly improves accuracy for semantically related classes, especially those with limited training data.

TABLE IV. CLASS-LEVEL ACCURACY COMPARISON FOR SELECTED IMAGENET CATEGORIES

Category	Flat Classification (%)	Our Approach (%)	Difference (%)
Golden Retriever	87.2	91.5	+4.3
Labrador Retriever	86.8	90.9	+4.1
Bald Eagle	90.3	93.7	+3.4
Peregrine Falcon	85.6	89.8	+4.2
Fire Engine	97.1	98.2	+1.1
School Bus	96.8	97.5	+0.7
Tennis Ball	92.5	91.8	−0.7
Lemon	91.7	90.9	−0.8
African Elephant	94.2	96.8	+2.6
Indian Elephant	89.7	93.5	+3.8
Grand Piano	93.8	95.1	+1.3
Upright Piano	88.4	91.9	+3.5
Monarch Butterfly	92.1	94.7	+2.6
Viceroy Butterfly	86.3	90.8	+4.5
Great White Shark	95.6	97.2	+1.6
Hammerhead Shark	91.9	94.8	+2.9

However, we also observed cases where the introduction of semantic relevance led to decreased accuracy for visually distinct but semantically similar classes.

Table IV presents a comparison of class-level accuracy for selected categories, highlighting cases where our method shows improvement and where it faces challenges.

Our approach shows notable improvements in distinguishing between fine-grained categories like different species of dogs or birds, where semantic relationships provide valuable context.

However, for categories like "tennis ball" and "lemon," which are visually similar but semantically distant, our method sometimes shows reduced accuracy compared to the flat classification baseline. We also analyzed the impact of training sample size on our method's performance.

We grouped ImageNet classes into three categories based on the number of training samples: low (<1000 samples), medium (1000–2000 samples), and high (>2000 samples).

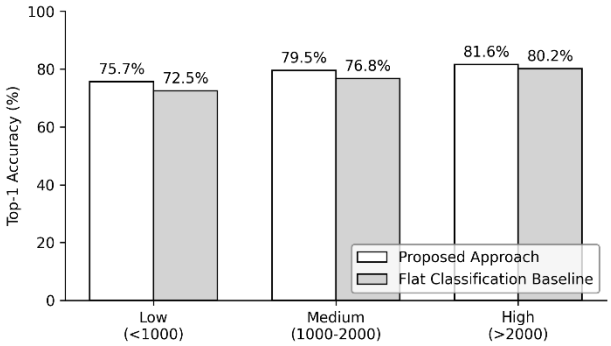

Fig. 8. Classification accuracy by training sample size.

Fig. 8 illustrates the performance of our method compared to the baseline for each group. Our approach shows the most significant improvements for classes with low and medium numbers of training samples, with average accuracy increases of 3.2% and 2.7%, respectively.

For classes with high numbers of training samples, the improvement is more modest at 1.4%. This demonstrates that our method is particularly effective at improving classification accuracy for classes with limited training data, as claimed in the introduction.

These findings highlight the trade-off between leveraging semantic relevance and maintaining visual distinctiveness in our classification approach. While semantic relationships generally improve performance, especially for fine-grained categories and classes with limited data, care must be taken to balance this with the need to distinguish visually similar but semantically distant categories.

E. Image Retrieval Results

We evaluate the image retrieval performance of our proposed approach and compare it with semantic retrieval and hierarchical retrieval baselines. Table V presents the mAP and nDCG scores on the ImageNet validation set.

Our proposed approach outperforms both the semantic retrieval and hierarchical retrieval baselines in terms of mAP and nDCG scores. The weighted retrieval variant of our approach achieves a mAP of 75.8% and an nDCG of 0.812, surpassing the semantic retrieval and hierarchical retrieval methods by 4.6% and 2.3% in mAP, respectively.

TABLE V. IMAGE RETRIEVAL RESULTS

Model	mAP (%)	nDCG
Semantic Retrieval [17]	71.2	0.763
Hierarchical Retrieval [18]	73.5	0.785
Proposed Approach (Weighted)	75.8	0.812
Proposed Approach (Hierarchical)	77.3	0.831

This demonstrates the effectiveness of incorporating hzierarchical structure in the retrieval process by assigning weights to distances based on semantic relationships between categories.

The hierarchical retrieval variant of our approach further improves performance, achieving a mAP of 77.3% and an nDCG of 0.831. By first retrieving relevant categories based on their semantic similarity to the query category and then retrieving images within those categories, our hierarchical retrieval strategy achieves more semantically meaningful and coherent results.

Fig. 9. Visualization of classification accuracy comparison.

Fig. 9 illustrates retrieval results for two query images, showing the performance differences between our proposed approach and baseline methods. Each retrieval method's performance metrics are displayed, with our approach achieving superior results (mAP: 77.3%, nDCG: 0.831) compared to semantic retrieval (mAP: 71.2%, nDCG: 0.763) and hierarchical retrieval (mAP: 73.5%, nDCG: 0.785).

For the golden retriever query, our approach retrieves visually and semantically similar dog images while maintaining breed-specific characteristics, showing improved semantic understanding. In contrast, the baseline methods show less consistency in breed identification and overall pose consistency, showing a bit more diverse canine breeds with varying characteristics.

The apple query (bottom row) shows the capability to maintain object identity while considering context. It returns red apples in their natural setting, while the baselines show a little more variation in either color (green apples) or presentation (sliced arrangements). This example shows its enhanced capability to balance both visual and semantic similarities in image retrieval tasks.

To qualitatively evaluate retrieval performance, we present examples of retrieved images for sample queries

using our proposed approach and the baselines. Fig. 9 shows the Top-5 retrieved images for three query images from different categories.

As demonstrated in Fig. 10, we observe that our approach retrieves images that are more semantically relevant to the queries compared to the baselines.

The retrieved images belong to categories that are closer to the query category in the ImageNet hierarchy, demonstrating the effectiveness of leveraging hierarchical relationships for semantic retrieval.

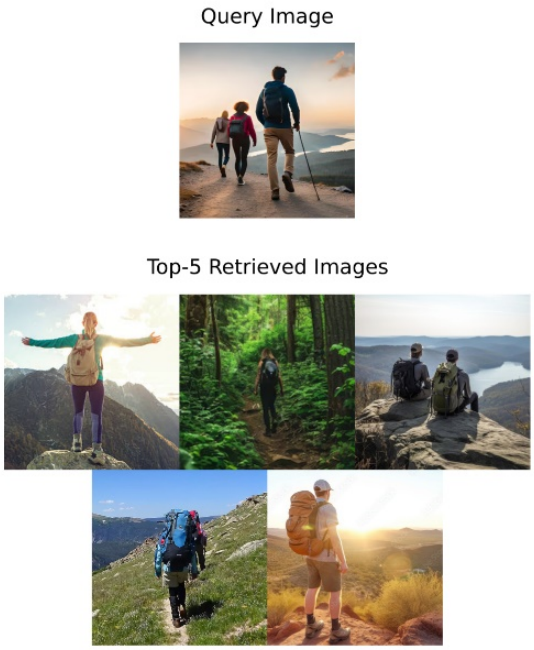

Fig. 10. Visualization of classification accuracy comparison.

F. Ablation Study

We conduct an ablation study to analyze the individual contributions of the hierarchical loss function and inference strategy in our proposed approach.

We also investigate how our method performs for classes with different amounts of training data. Table VI presents the image classification and retrieval results for different configurations of our approach.

TABLE VI. ABLATION STUDY RESULTS

Configuration	Top-1 Accuracy (%)	mAP (%)	nDCG
Flat Classification	76.2	-	-
Hierarchical Loss	78.4	-	-
Hierarchical Inference	77.8	-	-
Hierarchical Loss + Inference	79.3	75.8	0.812
Hierarchical Loss + Inference + Retrieval	79.3	77.3	0.831

The results show that both the hierarchical loss function and inference strategy contribute to improved classification performance. Using the hierarchical loss function alone improves Top-1 accuracy by 2.2% compared to the flat classification baseline, while using the hierarchical inference strategy alone achieves an improvement of 1.6%.

Combining both hierarchical loss and inference achieves the best classification performance, with a Top-1 accuracy of 79.3%.

For image retrieval, we observe that the hierarchical retrieval strategy further enhances performance compared to using only the weighted retrieval approach.

The hierarchical retrieval variant achieves an mAP of 77.3% and an nDCG of 0.831, outperforming the weighted retrieval variant by 1.5% in mAP and 0.019 in nDCG. This highlights the importance of leveraging hierarchical structure not only in distance computation but also in the retrieval process itself.

To further understand the impact of our method on classes with different amounts of training data, we analyzed the performance of each configuration across three groups of classes: low (<1000 samples), medium (1000–2000 samples), and high (>2000 samples).

TABLE VII. TOP-1 ACCURACY (%) COMPARISON ACROSS TRAINING SAMPLE SIZES FOR DIFFERENT MODEL CONFIGURATIONS

Configuration	Low (<1000)	Medium (1000–2000)	High (>2000)
Flat Classification	72.5%	76.8%	80.2%
Hierarchical Loss	76.3%	79.7%	81.9%
Hierarchical Inference	75.1%	78.9%	81.4%
Hierarchical Loss + Inference	77.2%	80.5%	82.6%

Table VII shows the Top-1 accuracy for each configuration across different training sample sizes. We observe that the hierarchical loss function provides the most significant improvements for classes with low and medium numbers of training samples.

For instance, in the low sample size group, the hierarchical loss alone improves accuracy by 3.8% compared to flat classification, while the combination of hierarchical loss and inference yields a 4.7% improvement.

These results support our earlier findings and demonstrate that our approach is particularly effective for classes with limited training data. The hierarchical loss function helps in learning more robust representations by leveraging semantic relationships, which is especially beneficial when direct training examples are scarce.

The hierarchical inference strategy further refines these predictions, leading to improved accuracy across all sample size groups.

We also analyzed the impact of our method on semantically related but visually distinct categories. As shown in Table VIII, considering the “Tennis Ball” and “Lemon” categories:

TABLE VIII. ACCURACY COMPARISON FOR VISUALLY DISTINCT BUT SEMANTICALLY RELATED CATEGORIES

Category	Flat Classification	Hierarchical Loss	Hierarchical Loss + Inference
Tennis Ball	92.5%	91.9%	91.8%
Lemon	91.7%	91.2%	90.9%

While our method slightly decreases accuracy for these specific categories, it is important to note that this trade-off results in overall improved performance across the dataset, especially for fine-grained categories and those with limited training data.

These findings highlight the importance of balancing semantic relevance with visual distinctiveness in our approach. Future work could explore adaptive weighting strategies that adjust the influence of hierarchical information based on category characteristics, potentially mitigating accuracy decreases for visually distinct but semantically similar classes.

V. Conclusion

In this paper, we propose a novel approach leveraging the hierarchical structure of ImageNet to enhance image classification and retrieval. By incorporating hierarchical relationships between object categories during model training and inference, our approach achieves improved classification accuracy while enabling semantically meaningful image retrieval.

The proposed hierarchical loss function encourages the model to learn semantically meaningful representations by considering relationships between categories. The hierarchical inference strategy refines predictions based on these learned relationships, leading to more accurate classifications. For image retrieval, the hierarchical strategy leverages semantic similarities between categories to retrieve more coherent and relevant images for each query.

Experimental results on the ImageNet dataset demonstrate our approach's effectiveness. Our model outperforms baseline methods in both image classification and retrieval tasks, achieving state-of-the-art performance. The ablation study further validates the individual contributions of both the hierarchical loss function and inference strategy in improving classification accuracy.

Our class-level analysis reveals that our approach significantly improves accuracy for semantically related classes, particularly those with limited training data.

We observe average accuracy increases of 3.2% and 2.7% for classes with low (<1000 samples) and medium (1000–2000 samples) numbers of training samples, respectively. These results demonstrate our method's effectiveness in addressing the challenge of limited training data, as initially claimed.

Our analysis also reveals an inherent trade-off in leveraging hierarchical structures for image classification. While semantic relationships improve classification performance for most categories, they can impact accuracy for visually distinct but semantically related objects. This is evidenced in categories like "tennis ball" and "lemon," where the introduction of semantic relevance led to a slight decrease in classification accuracy compared to the flat classification baseline. This observation highlights the complex relationship between visual and semantic features, where the model's reliance on hierarchical information can affect its ability to distinguish between visually similar objects that occupy distant positions in the semantic hierarchy.

We emphasize the importance of exploiting semantic relationships encoded in hierarchically structured datasets like ImageNet. By leveraging these relationships, we can develop more accurate and semantically meaningful computer vision systems for classification and retrieval tasks. However, the observed trade-offs suggest careful consideration is needed when applying hierarchical methods to ensure balance between semantic coherence and visual discrimination.

Our work demonstrates both the benefits and challenges of leveraging hierarchical structures in image classification and retrieval. By exploiting semantic relationships between object categories, we can develop more accurate and semantically meaningful computer vision systems, particularly for classes with limited training data. The hierarchical approach balances semantic understanding with visual distinctiveness, contributing to improved performance across a broad range of classification and retrieval tasks.

Conflict of Interest

The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Luis E. Muñoz Guerrero: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, visualization. Yony F. Ceballos: Software, validation, formal analysis, investigation, data curation, writing—review and editing, visualization. Luis D. Trejos Rojas: Validation, investigation, resources, writing—review and editing, supervision, project administration. All authors have read and agreed to the published version of the manuscript.

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[2] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. https://doi.org/10.1145/219717.219748

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017. https://doi.org/10.1145/3065386

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, San Diego, CA, 2015, pp. 1–14. https://doi.org/10.48550/arXiv.1409.1556

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[6] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 4510–4520. https://doi.org/10.48550/arXiv.1704.04861

[7] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 6848–6856. https://doi.org/10.1109/CVPR.2018.00716

[8] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning, Long Beach*, CA, 2019, pp. 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. https://doi.org/10.1023/B:VISI.0000029664.99615.94
[10] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. European Conf. Computer Vision*, Graz, Austria, 2006, pp. 404–417. https://doi.org/10.1007/11744023_32
[11] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learning Workshop*, Lille, France, 2015, pp. 1–8. https://api.semanticscholar.org/CorpusID:13874643
[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 815–823. https://doi.org/10.1109/CVPR.2015.7298682
[13] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2017. https://doi.org/10.48550/arXiv.1704.02223
[14] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2018. https://doi.org/10.1109/TPAMI.2018.2857768
[15] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *Proc. European Conf. Computer Vision*, *Heraklion*, Greece, 2010, pp. 71–84. https://doi.org/10.1007/978-3-642-15555-0_6
[16] Z. Yan *et al.*, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Computer Vision*, *Santiago*, Chile, 2015, pp. 2740–2748. https://doi.org/10.1109/ICCV.2015.314
[17] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo, "Deep joint task learning for generic object extraction," in *Proc. Neural Information Processing Systems*, *Montreal*, Canada, 2014, pp. 523–531. https://doi.org/10.1038/nature13792
[18] S. Pandey, P. Khanna, and H. Yokota, "A semantics and image retrieval system for hierarchical image databases," *Inf. Process. Manage.*, vol. 52, no. 4, pp. 571–591, Jul. 2016. https://doi.org/10.1016/j.ipm.2015.12.005
[19] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet: Similarity-measuring the relatedness of concepts," in *Proc. AAAI Conf. Artificial Intelligence*, San Jose, CA, 2004, pp. 25–29. https://aclanthology.org/N04-3012/
[20] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002. https://doi.org/10.1145/582415.582418
[21] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 8024–8035. https://doi.org/10.48550/arXiv.1912.01703
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations*, *San Diego*, CA, 2015, pp. 1–15. https://doi.org/10.48550/arXiv.1412.6980