Fine-Tuning EfficientNet Feature Extraction for Efficiency in Image Retrieval Applications

Ghazal Saadloonia[a], Ijaz Ahmad[b], and Seokjoo Shin[a]
[a]Dept. of Computer Engineering, Chosun University, Gwangju, Korea
[b]Dept. of Electrical and Computer Engineering, Korea University, Seoul, Korea
ghazal.saadloonia@chosun.ac.kr, ijaz@korea.ac.kr, sjshin@chosun.ac.kr (*Corresponding author*)

***Abstract*—Fine-tuning pre-trained classification models for image retrieval tasks can deal with the computational expense and data requirements of deep learning. However, standard classification networks are fundamentally designed to output probability distributions, which differs from the requirements of image retrieval tasks that rely on efficient and meaningful distance metrics in a feature space. Additionally, balancing a trade-off between semantic richness and efficient feature dimensioning is still a critical challenge. In this paper, we propose an adapted pre-trained EfficientNet models-based Content-Based Image Retrieval (CBIR) scheme designed to address these limitations by leveraging transfer learning and fine-tuning on the Corel-1K dataset. The proposed scheme incorporates critical post-processing steps, including feature normalization and the selection of an appropriate distance metric, and analyzes features extracted from intermediate layers to identify an optimal depth for a better trade-off between performance and computational efficiency. Through simulation analysis, we show that the proposed scheme outperforms established CBIR approaches. For example, among the analyzed models, EfficientNet-B4 model achieved the highest overall Mean Average Precision (mAP) score of 93.50%. Furthermore, our scheme maintains superior performance across all reported ranked metrics, indicating a more robust and effective feature representation for image retrieval tasks compared to prior techniques.**

***Index Terms*—EfficientNet, Corel-1K, transfer learning, content-based image retrieval.**

I. Introduction

Image retrieval systems have evolved significantly, moving from traditional keyword-based searches to methods that utilize the visual content of the images themselves, often known as Content-Based Image Retrieval (CBIR) systems. Early approaches relied heavily on handcrafted features such as color histograms, texture descriptors, and shape information [1], [2]. These methods offered interpretability and computational simplicity; however, they often struggle with complex, real-world image variability and semantic understanding [1]. The advent and rapid advancement of Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision, demonstrating state-of-the-art (SOTA) performance in feature extraction that can enable applications such as, image classification [1], [3].

Despite the powerful discriminative capabilities of classification networks, several challenges remain in applying them effectively to CBIR tasks. For example, high-performing CNN architectures are often computationally expensive and require large volume of data for their training. In this regard, transfer learning is a widely adopted paradigm that allows a model pre-trained on a large dataset (like ImageNet) to be used as a starting point for a new, related task, rather than training it from scratch. Consequently, this learned feature reuse reduces training time, the required data, and computational resources for the new, specific problem [4]. However, despite their overall effectiveness, pre-trained models often show performance limitations when transferred to different or domain-specific visual datasets as they are optimized for the source data statistical distribution. Therefore, they tend to extract fewer distinguishing features from classes that are underrepresented in the source domain [3]. Furthermore, standard classification networks are fundamentally designed to output probability distributions, which differs from the requirements of image retrieval tasks that rely on efficient and meaningful distance metrics in a feature space. Additionally, determining the optimal depth for feature extraction that balances rich semantic representation with efficient feature dimensioning remains a critical challenge.

In this paper, we propose an adapted pre-trained deep learning model-based CBIR scheme designed to address these limitations. Specifically, we leverage the EfficientNet models [5], known for their superior performance and efficient scaling properties, and fine-tune them using transfer learning on the Corel-1K dataset [6]. Although these models are originally designed for classification, we incorporate critical post-processing steps, including feature normalization and the selection of an appropriate distance metric compatible with the model's loss function, ensuring robust feature matching for optimal performance in distance-based image retrieval tasks. Furthermore, we analyze features extracted from intermediate layers to identify an application-specific optimal depth that provides a better trade-off between semantic representation and computational efficiency. Finally, we provide an extensive comparison with existing baseline CBIR methods and demonstrate superior retrieval performance in both overall and specific ranked metrics.

II. Proposed Method

Fig. 1 shows a high level illustration of our proposed CBIR scheme. In general, an image retrieval system consists of feature extraction and feature matching modules. This section outlines the complete pipeline for the proposed image retrieval system, detailing how we adapted a pre-trained deep

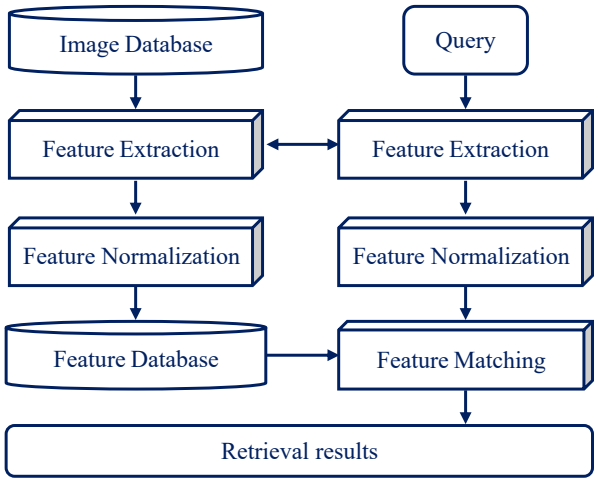

Fig. 1: Proposed CBIR scheme where the feature extraction is based on EfficientNet models.

learning model to serve as a robust feature extractor, the fine-tuning process, and the similarity matching procedure used for evaluation.

A. Feature Extraction for Image Retrieval

Feature extraction is a function that maps an input image I into a high-dimensional feature space Φ. In this work, we employ a pre-trained deep learning (DL) model and adapt it using transfer learning to perform this mapping effectively. The following sections detail the strategy for fine-tuning this model and generating robust feature representations.

1) Model Architecture and Transfer Learning: The DL model f used in this work is a parameterized function f_θ (based on the EfficientNetB1 architecture) that learns to update its parameters θ by minimizing a loss function L, such that the predicted output is as close as possible to the ground truth label. The most common choice for the loss function in classification tasks is the Categorical Cross-Entropy (CCE) loss, defined for a single sample as:

$$L_{\text{CCE}}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{1}$$

where C is the number of classes, y_c is a binary indicator if class c is the correct classification, and $\hat{y}_c$ is the model's predicted probability for class c.

The model is structured into two main components: a feature extractor module, consisting of multiple convolutional layers, and a classifier module, comprised of fully connected layers. A model's classifier module is highly specific to a certain task (for example, designed for 1000 ImageNet categories), while its underlying feature extractor module is generalizable. This modularity allows us to utilize the learned representations from one source task (T_1) to optimize the model efficiently for our target retrieval task (T_2), a process known as transfer learning.

The objective function for each task $\mathcal{T} \in \{T_1, T_2\}$ is defined as minimizing the expected loss:

$$\min_{\hat{F}_\mathcal{T} \in f_\mathcal{T}} L(\hat{F}_\mathcal{T}) = \min_{\hat{F}_\mathcal{T} \in f_\mathcal{T}} E\left[L_\mathcal{T}\left(Y_\mathcal{T}, \hat{F}_\mathcal{T}(X_\mathcal{T})\right)\right]. \tag{2}$$

The goal of transfer learning is to leverage the pre-trained model $\hat{F}_{T_1}$ from the source task to find the optimal estimator $\hat{F}_{T_2}$ for the target task T_2. This is achieved by retraining the classifier and fine-tuning specific layers or the entire feature extractor module.

2) Fine-Tuning Framework: The adaptation process follows a structured pipeline involving input and output transformations to align the source model with the target task requirements.

Step 1. Input Transformation: Input image in the target task must match the input specifications of the source task images. For which, we employ a transformation function $g_\mathcal{X}$ to resize images and normalize channels for the EfficientNet models.

Step 2. Applying the Pre-trained Model: The transformed input data is passed through the pre-trained model $\hat{F}_{T_1}$, defined as:

$$\hat{F}_{T_1}(g_\mathcal{X}(X_{T_2})) \in \mathcal{Y}_{T_1}. \tag{3}$$

Step 3. Output Transformation and Fine-Tuning: The source classifier layers are discarded because they are optimized for the source task's specific 1000 categories and are not relevant to our target retrieval task. The disparity between the source and target output spaces necessitates an output transformation $g_\mathcal{Y}$. This function maps the output from Step 2 into the target output space $\mathcal{Y}_{T_2}$. This involves attaching new layers (for example, a global pooling layer followed by a dense layer) and fine-tuning the entire network using the target objective L_{T_2}.

3) Model Truncation and Embedding Generation: Following the fine-tuning process, the network is repurposed for retrieval by extracting the intermediate representation. The following steps describe how the final feature embeddings are generated.

Model Truncation: The fine-tuned model $\hat{F}_{T_2}$ is truncated by removing the final task-specific classification layer, which isolates the feature extractor module from the final decision-making mechanism.

Feature Embedding: The modified network acts as a fixed feature extractor function ϕ, mapping an input image I to a high-dimensional feature vector $\mathbf{v} \in \mathbb{R}^D$ (where D is the output dimension of the feature embeddings): $\phi(I) = \mathbf{v}$. Using this feature extractor ϕ, the feature vectors s and q are respectively obtained from the stored and query images.

L2 Normalization: Features are L2 normalized before similarity matching to ensure all embeddings lie on a unit hypersphere:

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}. \tag{4}$$

This normalization step is crucial, as it prepares the embeddings for robust angular comparison in the subsequent step.

(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 2: Example images from the Corel-1K dataset.

B. Feature Similarity Measurement

The final step of the retrieval process involves the feature matching module. For which, we employ a distance metric to quantify the similarity between a query feature vector q and all stored feature vectors s. The efficacy of this choice depends heavily on how the embeddings were optimized during training. This work uses Cosine similarity as the primary metric, which measures the angle between two normalized vectors in the high-dimensional space:

$$\text{Distance}(q, s) = 1 - \frac{q \cdot s}{\|q\|\|s\|}. \tag{5}$$

The choice of Cosine similarity is particularly compatible with the cross-entropy loss function (CCE) used during training. The CCE optimizes primarily for the direction of the feature vectors rather than their magnitude. By applying L2 normalization inherent to Cosine similarity calculation, we isolate this angular information. In contrast, the L2 (Euclidean) distance is sensitive to both direction and magnitude, making it less suitable for the variance in vector lengths typically produced by CCE training. Aligning the retrieval metric with the training objective ensures robust performance.

III. Simulation Results

A. Dataset

The Corel-1K dataset [6] is utilized as the benchmark dataset for evaluating the performance of our proposed image retrieval methodology. It consists of a total 1,000 images organized into 10 distinct categories, with each category containing exactly 100 images. The images have consistent dimensions of either 384×256 or 256×384 pixels. The 10 categories included in the dataset are diverse and cover a variety of natural and man-made scenes. Specifically, the "Dinosaur" category is unique within this dataset as it contains synthetic illustrations rather than natural photographs, which presents a specific challenge for models pre-trained on photographic content like ImageNet. For visual analysis, Fig. 2 shows an example image for each label in the this dataset.

B. Experimental Setup

For analysis, we implemented our proposed Content-Based Image Retrieval (CBIR) scheme using pre-trained EfficientNet models [5], specifically configurations B0 through B5. These models were pre-trained on the large-scale ImageNet dataset to leverage transfer learning capabilities. The Corel-1K dataset was used for training, fine-tuning, and evaluation. For reliable performance assessment, the dataset was divided into an 80% training set and a 20% testing set. All experiments were repeated three times using different random seeds for the dataset splits to ensure robustness. The models were optimized using the Adam optimizer with categorical cross-entropy as the loss function and a batch size of 8. The training process was conducted in two distinct phases:

Phase I: Classifier Training. Initially, only the top classification layers of the EfficientNet models were trained. The base layers were kept frozen. This phase ran for 5 epochs using a learning rate of 0.001.

Phase II: Fine-tuning. Following Phase I, we unfroze the top 80 layers of the entire model and then fine-tuned for an additional 15 epochs using a reduced learning rate of $1\text{e}-5$ to ensure stable convergence and prevent catastrophic forgetting of the pre-trained weights.

C. Evaluation Metrics

To rigorously assess the performance of the image retrieval system, we utilize several widely accepted metrics that provide a comprehensive view of ranking quality and efficiency.

1) Average Precision (AP) and Mean Average Precision (mAP): This work uses the Mean Average Precision (mAP) as the primary metric, which is derived from the Average Precision (AP) score. The AP metric measures the retrieval performance in terms of how well relevant images are ranked within the list of retrieved images for a single query. For a single query q, the AP is defined as:

$$\text{AP}(q) = \sum_{k=1}^{N} P(k) \times \text{rel}(k), \tag{6}$$

where N is the total number of images in the database, $P(K)$ is the precision at rank k, and $\text{rel}(k)$ is a binary relevance indicator (1 if the item at rank k is relevant, 0 otherwise). From (6), the mAP score is derived as the average of the AP scores calculated across all Q queries in the test set:

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}(q). \tag{7}$$

2) Precision @K (P@K) and Recall @K (R@K): P@K and R@K are used to evaluate the quality of the top K retrieved results, which directly correlates with the user's immediate experience. P@K measures the retrieval performance as how many retrieved images are relevant among the top K results:

$$\text{P@K} = \frac{|\mathcal{D}^l \cap \mathcal{D}^r|}{|\mathcal{D}^r|}, \tag{8}$$

where $\mathcal{D}^l$ and $\mathcal{D}^r$ respectively denotes the number of relevant and retrieved images. R@K measures the retrieval performance as how many relevant images are correctly identified

within the top K results relative to the total number of relevant images in the database:

$$\text{R@K} = \frac{|\mathcal{D}^l \cap \mathcal{D}^r|}{|\mathcal{D}^l|}. \tag{9}$$

These metrics provide a balanced assessment of the image retrieval system's ability to find relevant items quickly and comprehensively.

D. Retrieval Performance Analysis

In this section, we conduct a thorough evaluation of the proposed fine-tuned EfficientNet architectures using the Corel-1K dataset. The analysis is structured into three main parts: first, an investigation into the optimal feature extraction depth; second, a performance comparison against existing state-of-the-art retrieval techniques; and finally, a detailed class-wise performance breakdown to identify category-specific strengths and weaknesses.

Analysis of EfficientNet Feature Extraction Points. Table I summarizes the extensive evaluation of fine-tuned EfficientNet models (B0 through B5) for image retrieval on the Corel-1K dataset. The comparison focuses on feature extraction from two distinct network depths that is, Block 6 versus Block 7, chosen to investigate a potential optimal point that balances rich semantic representation with efficient feature dimensioning. The values are reported as the mean $\pm$ standard deviation across three independent experimental runs for mAP, mAP@K, P@K, and R@K metrics. It is noteworthy that standard mAP evaluates the quality across the entire ranked list of retrieved images, while @K metrics specifically focus on performance within the top K results. The consistently low standard deviation values indicate high stability and reproducibility of all results.

A primary observation across all tested models is that features extracted from Block 7 consistently yield superior retrieval performance in terms of overall mAP and metrics reliant on deeper result lists (for example, P@100, R@100), validating the identification of this layer as an optimal point for richer semantic representation. However, a focused analysis of the ranked metrics (K=10 or 20) reveals that the performance difference between Block 6 and Block 7 is not statistically significant in the very early results. These metrics are highly relevant for practical applications where only the top few results are immediately presented to the user. This finding highlights a crucial trade-off that Block 6 provides smaller feature vectors (for example, d=192 vs. d=320 for B0), which facilitates significantly faster feature matching and reduced computational overhead. Therefore, Block 6 is a viable alternative for applications prioritizing real-time performance and faster response times over marginal gains in the overall result list quality. Furthermore, simply scaling up the model size (from B0 to B5) does not guarantee a proportional increase in performance for example, the B5 model, despite having the largest feature dimensions, does not achieve peak efficacy.

In summary, the EfficientNet-B4 model with Block 7 features emerged as the overall top performer across the majority of metrics presented in the table. For subsequent analysis, two models were selected to balance performance and computational efficiency: the B4 Block 7 configuration (for best performance) and the B0 Block 7 configuration (for better computational complexity).

Comparison with the State-of-the-Art Techniques. Table II presents a comparison of the proposed EfficientNet-B0 and B4 (Block 7) models against several established image retrieval techniques from existing literature, evaluated on the Corel-1K dataset. The baseline methods include early deep learning approaches such as AlexNet [7] and VGG16 [8], along with custom CNN architectures [9], and hybrid methods combining handcrafted features (for example, LBP, HOG, color histograms) with machine learning techniques like genetic algorithms for feature selection [10]. Note that due to variations in evaluation methodologies across studies, not all metrics are available for all baselines.

The proposed models demonstrate highly competitive performance, frequently establishing a new state-of-the-art on this dataset. For example, our B4 model achieves the highest overall mAP observed in the comparison (93.50%), significantly outperforming both the robust fusion technique of [7] and their basic AlexNet. We also observe superior performance in ranked metrics; our B4 model's mAP@10 surpasses the custom CNN of [9]. Furthermore, a key strength of our approach is visible in the precision and recall metrics across varying K values. Although several baselines achieve high P@K scores in specific instances (for example, [8]'s Model II achieves 96.25% P@20), our B0 and B4 models maintain high performance across all reported ranked metrics (P@10, P@20, P@100), R@K), indicating superior consistency and depth in retrieval quality.

The performance gap is particularly evident in the R@100 metric, where our B4 model achieves 90.50%, drastically outperforming hybrid methods like [10] (62.65%) and [8]'s Model I (32.2%). This highlights the superior ability of the fine-tuned EfficientNet features to retrieve a higher proportion of the total relevant images within the database. The results strongly suggest that the features extracted from the optimal point (Block 7) of the EfficientNet architecture provide a more robust and effective representation for image retrieval tasks compared to prior techniques evaluated on this dataset.

Class-Wise Retrieval Performance Analysis. Table III provides a detailed class-wise comparison of retrieval performance (mAP@10, mAP@20, and mAP@100) between our proposed EfficientNet models (B0 and B4, Block 7) and the hybrid approach of [10]. As [10] was the only existing scheme that reported performance metrics on a per-class basis on the Corel-1K dataset, it serves as our specific baseline for this granular analysis.

The results reveal that the proposed models generally outperform the baseline across most classes and all evaluated metrics. The average performance across the 10 classes (bottom 'Mean' row) clearly favors our approach, with the B4 model achieving the highest overall mean mAP@100 score

TABLE I: Performance analysis of fine-tuned EfficientNet models on Corel-1K dataset. The metric values are reported as mean±standard deviation. The best values are in **bold**.

Model	Block	d	mAP	mAP@10	mAP@20	mAP@100	P@10	P@20	P@100	R@10	R@20	R@100
B0	Block6	192	0.869±0.007	0.956±0.001	0.937±0.002	0.787±0.008	0.952±0.021	0.949±0.002	0.821±0.007	0.096±0	0.19±0	0.821±0.007
	block7	320	0.913±0.005	0.97±0.001	0.96±0.003	0.85±0.008	0.975±0.001	0.969±0.002	0.874±0.006	**0.098±0**	0.194±0	0.874±0.006
B1	Block6	192	0.851±0.003	0.953±0.002	0.934±0.002	0.767±0.003	0.961±0.002	0.947±0.002	0.802±0.003	0.096±0	0.189±0	0.802±0.003
	block7	320	0.921±0.007	0.975±0.002	0.966±0.003	0.868±0.008	0.979±0.002	0.972±0.002	0.887±0.007	**0.098±0**	0.194±0	0.887± 0.007
B2	block6	208	0.866±0.005	0.959±0.001	0.942±0.002	0.781±0.006	0.966±0.001	0.954±0.002	0.816±0.005	**0.097±0**	0.191±0	0.816±0.005
	block7	352	0.934±0.009	0.981±0.002	0.973±0.003	0.884±0.013	0.984±0.001	0.979±0.002	0.9±0.011	**0.098±0**	**0.196±0**	0.9±0.011
B3	block6	232	0.876±0.002	0.963±0.001	0.946±0.001	0.798±0.002	0.97±0.001	0.957±0.001	0.83±0.002	**0.097±0**	0.191±0	0.83±0.002
	block7	384	**0.938±0.005**	0.98±0.002	0.974±0.002	**0.892±0.007**	0.984±0.001	0.98±0.002	**0.907±0.006**	**0.098±0**	**0.196±0**	**0.907±0.006**
B4	block6	272	**0.877±0.004**	**0.967±0.001**	**0.951±0.002**	**0.804±0.006**	**0.973±0.001**	**0.961±0.002**	**0.834±0.005**	**0.097±0**	**0.192±0**	**0.834±0.005**
	block7	448	0.935±0.007	**0.981±0**	**0.974±0.001**	0.891±0.007	**0.984±0**	**0.98±0.001**	0.905±0.007	**0.098±0**	**0.196±0**	0.905±0.007
B5	block6	304	0.849±0.002	0.96±0.001	0.938±0.001	0.765±0.002	0.967±0.001	0.95±0.001	0.802±0.002	**0.097±0**	0.19±0	0.802±0.002
	block7	512	0.922±0.003	0.978±0.001	0.969±0.001	0.875±0.005	0.981±0.001	0.975±0.001	0.891±0.004	**0.098±0**	0.195±0	0.891± 0.004

TABLE II: Benchmarking of proposed EfficientNet models against state-of-the-art image retrieval techniques on Corel-1K dataset. The best values are in **bold**.

Method	mAP	mAP@5	mAP@10	mAP@20	mAP@100	P@10	P@20	P@100	R@10	R@20	R@100
[7]'s AlexNet	75.48	93.14	91.87	-	-	-	-	-	-	-	-
[7]'s	91.65	96.02	95.80	-	-	-	-	-	-	-	-
[8]'s VGG16	-	-	-	-	-	-	94.60	-	-	18.92	-
[8]'s Model I	-	-	-	-	-	95.2	87.25	32.2	9.52	17.45	32.2
[8]'s Model II	-	-	-	-	-	-	96.25	-	-	19.25	-
[9]	-	-	95.62	-	-	-	-	-	-	-	-
[10]	-	-	95.55	93.90	82.51	93.07	87.13	62.65	9.31	17.43	62.65
Ours (B0 Block7)	91.30	97.80	97.00	95.98	85.04	97.50	96.90	87.40	**9.80**	19.40	87.40
Ours (B4 Block7)	**93.50**	**98.65**	**98.10**	**97.43**	**89.10**	**98.40**	**98.00**	**90.50**	**9.80**	**19.60**	**90.50**

TABLE III: Per-class retrieval performance (mAP@K) comparison with the baseline method of [10]. The best values are in **bold**.

Labels	mAP@10			mAP@20			mAP@100		
	[10]	B0 (Ours)	B4 (Ours)	[10]	B0 (Ours)	B4 (Ours)	[10]	B0 (Ours)	B4 (Ours)
Africans	95.51	91.53	**96.21**	94.31	88.48	**95.17**	**83.81**	63.38	78.96
Architecture	92.40	**100.00**	**100.00**	89.86	**100.00**	**100.00**	73.48	92.01	**98.50**
Beach	86.31	**100.00**	**100.00**	82.13	**100.00**	**100.00**	63.28	98.23	**99.78**
Buses	**99.30**	97.98	98.25	**98.64**	96.74	97.89	**92.96**	84.99	86.07
Dinosaur	**100.00**	88.72	89.94	**100.00**	85.99	86.46	**99.96**	60.66	61.90
Elephant	97.27	**100.00**	**100.00**	94.88	**100.00**	**100.00**	77.36	**99.84**	99.81
Food	95.28	95.1	98.01	93.87	93.19	**96.69**	**81.84**	73.69	78.06
Horses	99.74	**100.00**	**100.00**	99.16	**100.00**	**100.00**	90.53	98.28	**99.58**
Mountains	89.85	**100.00**	**100.00**	86.59	**100.00**	99.94	68.33	97.04	**98.38**
Roses	**99.89**	96.24	98.53	**99.58**	95.43	98.12	**93.53**	82.26	90.01
Mean	95.55	96.96	**98.10**	93.90	95.98	**97.43**	82.51	85.04	**89.10**

of 89.10%, significantly higher than [10]'s 82.51%. Also, in several categories, particularly "Architecture," "Beach," "Elephant," and "Mountains," our models achieve near-perfect scores (100% mAP@10 and mAP@20), demonstrating superior ability in retrieving images within these specific semantic categories compared to the baseline.

However, the comparison also highlights specific weaknesses. For instance, in the "Africans" category, the baseline [10] maintains strong early precision (95.51% mAP@10) where our B0 model dips (91.53%). A notable observation is the "Dinosaur" class. Here, the hybrid method of [10] achieves perfect mAP@10 and mAP@20 scores, while our models show reduced performance in early retrieval metrics. This discrepancy is likely attributable to the nature of the "Dinosaur" images within the Corel-1K dataset, which are synthetic illustrations. Since our models are fine-tuned from a network pre-trained exclusively on natural images (ImageNet), their features are optimized for photographic content. The superior performance of [10] in this specific instance suggests that their handcrafted features are better suited for distinguishing highly specific, non-photographic visual patterns present in that unique class. Fig. 3 and Fig. 4 respectively show an example image from the "Dinosaur" and "Roses" classes along with retrieval results for our proposed EfficientNet models (B0 and B4, Block 7). Fig. 3 shows that for the underrepresented class of "Dinosaur" the models relied on color information as appeared in the wrong retrieval results.

Overall, despite minor performance dips in one or two specific classes driven by data distribution differences, the general superiority of our B4 model across all ranked metrics and across the majority of semantic categories is confirmed by this class-wise analysis.

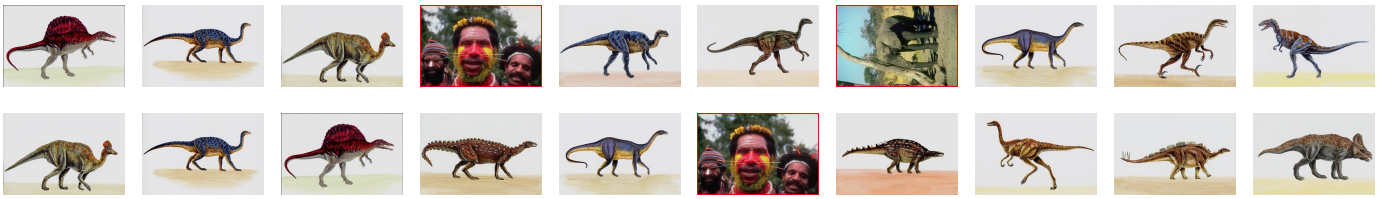

Fig. 3: Visual analysis of retrieving relevant images of "Dinosaur" class to a query image from the database using proposed CBIR scheme. Results for EfficientNet-B0 Block 7 is in the top row and for EfficientNet-B4 Block 7 is in the bottom row. The wrong retrieval results are surrounded in red box.

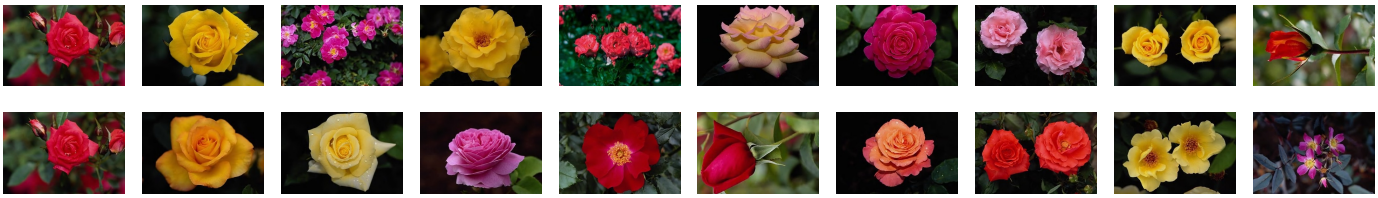

Fig. 4: Visual analysis of retrieving relevant images of "Roses" class to a query image from the database using proposed CBIR scheme. Results for EfficientNet-B0 Block 7 is in the top row and for EfficientNet-B4 Block 7 is in the bottom row.

IV. Conclusion

This study addressed critical challenges in applying pre-trained classification models effectively to Content-Based Image Retrieval (CBIR) tasks, focusing on optimizing efficiency, data requirements, and metric (classification model loss function and feature matching distance metric) compatibility. For this purpose, we proposed and evaluated an adapted pre-trained EfficientNet-based CBIR scheme that incorporated critical post-processing steps and analyzed features from intermediate layers to optimize performance trade-offs. Extensive evaluation of the Corel-1K dataset confirmed the scheme's efficacy and established a new benchmark in retrieval performance compared to existing schemes. Furthermore, a primary finding was the consistent superiority of features extracted from deeper layers (specifically Block 7 of EfficientNet-B4) for overall result list quality. Crucially, the analysis of ranked metrics highlighted a significant practical trade-off that using features from the shallower Block 6 offered comparable top-ranked performance while providing smaller feature vectors and faster computational speeds. In the future, we are interested in exploring dimensionality reduction algorithms for better feature matching efficiency.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (RS-2023-00278294).

References

[1] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2022.

[2] C. Sharma, G. Poornalatha, and K. B. Ajitha Shenoy, "A comprehensive review of recent advances in multimodal multimedia indexing and retrieval," *IEEE Access*, vol. 13, pp. 143 688–143 712, 2025.

[3] G. Saadloonia, I. Ahmad, and S. Shin, "Efficientnet models with dimensionality reduction for image retrieval applications," in *The 6th Korea Artificial Intelligence Conference*. KICS, 2025, pp. 243–244.

[4] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[6] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, 2001.

[7] A. Shakarami and H. Tarrah, "An efficient image descriptor for image classification and CBIR," *Optik*, vol. 214, p. 164833, 2020.

[8] J. Pradhan and H. Nenavath, "DNA-CBIR: DNA translation inspired codon pattern-based deep image feature extraction for content-based image retrieval," *IEEE Transactions on NanoBioscience*, vol. 24, no. 3, pp. 318–330, 2025.

[9] H. Rastegar and D. Giveki, "Designing a new deep convolutional neural network for content-based image retrieval with relevance feedback," *Computers and Electrical Engineering*, vol. 106, p. 108593, 2023.

[10] A. Roodaki, M. Sotoodeh, and M. Reza Moosavi, "Genetic algorithm-based feature selection from high-dimensional descriptors for improved content-based image retrieval," *IEEE Access*, vol. 13, pp. 198 034–198 053, 2025.