Implementation of Convolutional Neural Networks for Eyeglass Product Image Retrieval: A Comparative Study of ResNet-50 and MobileNetV2

Handri Taufik[1*], Sajarwo Anggai[2], Taswanda Taryo[3]
Master's Program in Informatics Engineering, Postgraduate Program, Pamulang University, South Tangerang
Jalan Raya Puspiptek No. 46, Buaran, Kecamatan Serpong, Kota Tangerang Selatan, Banten 15310
Email: handrith@gmail.com[1], sajarwo@ gmail.com[2], taswandataryo@ gmail.com[3]

The increasing similarity among eyewear product designs poses significant challenges for conventional text-based search systems, highlighting the need for effective Content-Based Image Retrieval (CBIR) approaches. This study proposes a CNN-based CBIR system for eyeglass frame and sunglasses retrieval, employing a comparative analysis of ResNet50 and MobileNetV2 as feature extractors. The dataset comprises 4,500 gallery images and 300 query images, with feature similarity measured using cosine similarity and accelerated through FAISS indexing. Experimental results indicate that ResNet50 achieves higher recall (0.0622), demonstrating its ability to capture more complex visual features. In contrast, MobileNetV2 provides superior ranking performance, achieving an mAP of 0.6091 and an MRR of 0.1427, outperforming ResNet50 (mAP of 0.5019 and MRR of 0.0713), while also reducing feature extraction time (0.1348 s versus 0.2023 s). These findings suggest that ResNet50 is more suitable for accuracy-oriented retrieval tasks, whereas MobileNetV2 is better suited for real-time and resource-constrained applications.
Keywords: CBIR, CNN, ResNet50, MobileNetV2, Cosine Similarity.

Corresponding Author:
Handri Taufik
Master's Program in Informatics Engineering, Postgraduate Program, Pamulang University, South Tangerang
Jalan Raya Puspiptek No. 46, Buaran, Kecamatan Serpong, Kota Tangerang Selatan, Banten 15310
handrith@gmail.com

1. Introduction

The rapid growth of the optical industry has resulted in a substantial increase in the number and diversity of frame and sunglasses products, many of which exhibit highly similar visual characteristics. This proliferation of visually homogeneous products creates significant challenges for product search, identification, and differentiation in digital catalogs and e-commerce platforms. Conventional retrieval systems predominantly rely on text-based metadata such as brand, model, material, and color, which are often incomplete, subjective, or inconsistently annotated across sellers and platforms. As a consequence, a semantic gap emerges between user intent and retrieved results, leading to suboptimal retrieval performance and a degraded user experience[1]. These limitations are particularly pronounced in fashion and eyewear domains, where subtle variations in shape, curvature, and design details play a critical role in consumer decision-making[2].

To overcome the inherent weaknesses of text-based retrieval, Content-Based Image Retrieval (CBIR) has been proposed as an alternative paradigm that enables image search based on intrinsic visual features, including shape, color distribution, and texture patterns. CBIR reduces dependence on manually curated textual annotations by directly exploiting visual representations extracted from images, thereby offering a more natural and intuitive retrieval mechanism for visually driven product search scenarios[3][4]. The emergence of deep learning has further transformed CBIR, with Convolutional Neural Networks (CNNs) becoming the dominant approach for learning discriminative and hierarchical feature embeddings that

capture both low-level visual cues and high-level semantic concepts[5][6]. Compared to traditional hand-crafted descriptors, CNN-based features have demonstrated superior robustness and generalization capabilities across diverse image retrieval tasks, including instance retrieval and fine-grained product matching[1].

Among the CNN architectures commonly employed as feature extractors, ResNet50 and MobileNetV2 represent two contrasting design philosophies in terms of model depth and computational efficiency. ResNet50 leverages deep residual connections to facilitate the training of very deep networks, enabling the extraction of complex and fine-grained visual representations that are effective for capturing subtle inter-class and intra-class variations[7][8]. In contrast, MobileNetV2 is specifically designed for efficiency, employing depthwise separable convolutions and inverted residual bottleneck structures to significantly reduce model parameters and computational cost while maintaining competitive representational power [9]. Prior studies have shown that lightweight CNN architectures can achieve performance comparable to deeper networks in various visual recognition tasks, particularly when computational resources and latency constraints are critical considerations[10][11][12]. However, most existing comparative studies primarily focus on classification accuracy rather than ranking-oriented retrieval performance, which is more relevant for CBIR systems deployed in practical product search environments[13][14][15].

Furthermore, CBIR performance is not solely determined by feature extraction quality but also by the effectiveness of similarity measurement and large-scale indexing mechanisms. Efficient similarity search frameworks, such as FAISS, have been introduced to enable scalable nearest-neighbor retrieval over high-dimensional CNN feature spaces, thereby supporting real-time retrieval even in large product catalogs[16]. In this context, the practical deployment of CBIR systems in e-commerce settings necessitates a careful balance between retrieval accuracy, ranking quality (e.g., top-ranked relevance), and computational efficiency, particularly when operating under resource-constrained environments such as mobile devices or edge computing infrastructures[17].

Despite the extensive adoption of ResNet-based and MobileNet-based architectures in computer vision applications, comparative analyses that systematically evaluate their trade-offs in terms of ranking quality (e.g., mAP, MRR), retrieval coverage (recall), and operational efficiency (inference time) for visually similar eyewear products remain limited. This gap is especially relevant for digital optical catalogs, where user satisfaction is strongly influenced by the relevance of top-ranked results and system responsiveness. Accordingly, this study designs and evaluates a CNN-based CBIR system for frame and sunglasses retrieval by conducting a systematic comparison of ResNet50 and MobileNetV2 as feature extractors. The evaluation employs multiple retrieval and efficiency metrics, including Precision@K, Mean Average Precision (mAP), Mean Reciprocal Rank (MRR), recall, and feature extraction time, to provide a comprehensive assessment of the accuracy–efficiency trade-offs between deep and lightweight CNN architectures. The findings are expected to offer empirical insights to guide practitioners and system designers in selecting appropriate CNN backbones for CBIR deployment in digital optical catalogs and e-commerce platforms, where both ranking quality and real-time performance are critical requirements.

2. Literature Review and Problem Statement

Recent advances in Content-Based Image Retrieval (CBIR) have shown that deep convolutional features substantially outperform traditional hand-crafted descriptors in instance-level and fine-grained retrieval tasks. Deep learning enables the extraction of hierarchical visual representations that capture both low-level textures and high-level semantic patterns, leading to more discriminative embeddings for similarity-based retrieval[3][6]. Residual networks, particularly ResNet architectures, have been widely adopted as feature extractors due to their ability to learn deep representations through residual connections, which

mitigate vanishing gradient problems and enhance feature expressiveness[7]. At the same time, the growing demand for real-time and large-scale deployment has motivated the development of lightweight CNN architectures such as MobileNetV2, which employ depthwise separable convolutions and inverted residual bottlenecks to achieve competitive performance with substantially reduced computational cost[9]. Prior studies indicate that lightweight CNNs can achieve accuracy comparable to deeper networks in various visual recognition tasks, especially under resource constraints[10][11]. In parallel, scalable similarity search frameworks such as FAISS have become essential components of CBIR pipelines, enabling efficient nearest-neighbor retrieval over high-dimensional CNN feature spaces in large image collections and supporting practical deployment in real-world systems[16].

Despite these methodological advances, existing comparative studies predominantly emphasize classification accuracy or generic image retrieval benchmarks, with limited focus on ranking-oriented retrieval performance and operational efficiency in product-specific domains characterized by high visual similarity, such as eyewear catalogs. In practical e-commerce environments, user satisfaction is largely determined by the relevance of top-ranked retrieval results (e.g., Mean Average Precision and Mean Reciprocal Rank) and system responsiveness, rather than by coverage alone. Consequently, a research gap persists in systematically examining the trade-offs between deep residual networks and lightweight CNN architectures in CBIR systems for visually homogeneous product categories, where subtle geometric and stylistic variations are critical for product differentiation. Based on this gap, the research problem of the present study is formulated as follows: *How do ResNet50 and MobileNetV2 differ in terms of ranking quality, retrieval effectiveness, and computational efficiency when deployed as feature extractors in a CNN-based CBIR system for frame and sunglasses product retrieval?* Addressing this problem, the study aims to provide empirically grounded guidance for selecting CNN backbones that balance accuracy and efficiency in the implementation of CBIR systems for digital optical catalogs and e-commerce platforms.

3. Method

This study employs an experimental approach aimed at evaluating the performance of a CNN-based CBIR system in retrieving frame and sunglasses products based on visual image similarity. The research workflow includes data collection, image preprocessing, feature extraction, similarity measurement, indexing, and system performance evaluation, as illustrated in Figure 1.

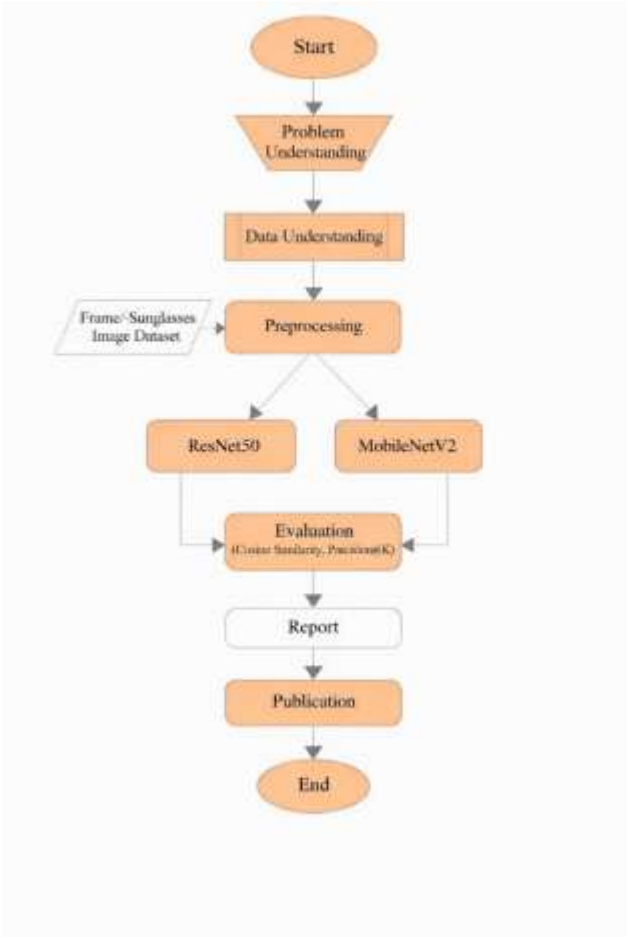

Figure 1. Research Work Procedure

The dataset used in this study consists of frame and sunglasses product images collected from a digital catalog, comprising more than 4,500 images as gallery data and 300 images as query data. All images are preprocessed through resizing and normalization to meet the input requirements of the CNN models. The dataset is designed to represent variations in product design, color, and shape commonly found in the optical industry.

Feature extraction is performed using two CNN architectures, namely ResNet50 and MobileNetV2, which are pretrained on the ImageNet dataset. The classification layers of each model are removed so that the extracted features are obtained from the final convolutional layers as visual representations of the images. The resulting feature vectors serve as the basis for similarity-based image retrieval. Image similarity is measured using cosine similarity, which evaluates the angular closeness between feature vectors. To improve retrieval efficiency, the gallery feature vectors are indexed using FAISS (Facebook AI Similarity Search), enabling fast nearest neighbor search even with large-scale data.

System performance is evaluated using several metrics, including Precision@10, Average Precision (AP@10), Mean Average Precision (mAP), Mean Reciprocal Rank (MRR), as well as feature extraction time. These metrics are employed to assess retrieval quality, ranking accuracy, and computational efficiency of the evaluated CNN models.

4. Results and Discussion

At the Data Understanding stage, the research dataset consists of images of frame and sunglasses products obtained from a digital optical catalog, where historical product images are used as the retrieval database and newly added product images serve as query inputs.

Figure 2. Eyeglass Product Dataset

The data preparation stage consists of more than 4,500 images, which are divided into 300 test images and 4,200 training images, as illustrated in Figure 2. All images undergo a preprocessing stage that includes resizing to 224×224 pixels and normalization according to the ImageNet standard. Both CNN models are employed as feature extractors using a transfer learning approach, in which the final classification layers are removed. Feature representations are extracted from the Global Average Pooling layer, producing 2048-dimensional feature vectors for ResNet50 and 1280-dimensional feature vectors for MobileNetV2. These feature vectors semantically represent the visual characteristics of the products.

Similarity matching is performed using cosine similarity to measure the closeness between query images and images in the database. The system then returns the top-K most visually relevant products. System performance is evaluated from two main aspects: effectiveness and efficiency. Effectiveness is measured using retrieval metrics such as Precision@K and Mean Average Precision (mAP), while efficiency is analyzed based on computational time and memory consumption for each model. This methodology enables an analysis of the trade-off between accuracy and efficiency, providing an empirical basis for selecting an appropriate CNN architecture for practical and scalable CBIR implementation in the optical retail domain.

The testing phase is conducted using two approaches: single testing and batch testing. In the single test scenario, a single image is captured or selected and then compared with the training dataset to measure similarity, with the system displaying the top-5 results along with the feature extraction time. Adjustable hyperparameters include the similarity threshold and the application of image augmentation to the query image. Two testing conditions are applied: without augmentation and with augmentation. The results obtained from the single test without augmentation are as follows:

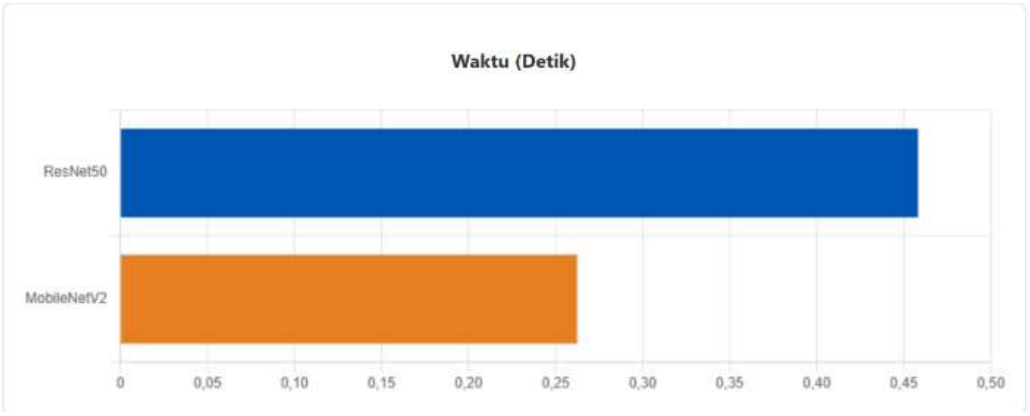

Figure 3. Non-TTA Time Comparison

Based on Figure 3, it can be observed that the feature extraction time of MobileNetV2 (0.2630 seconds) is significantly faster than that of ResNet50 (0.4589 seconds). The magnitude of this difference can be quantified through a simple mathematical analysis:

a. Time difference = 0.4589 – 0.2630 = 0.1959 seconds
b. Speed ratio = 0.4589 / 0.2630 = 1.74, indicating that MobileNetV2 operates approximately 1.74 times faster than ResNet50. In other words, within the time required for ResNet50 to process one image, MobileNetV2 is able to process nearly two images.
c. Percentage of computational savings = ((0.4589 – 0.2630) / 0.4589) × 100% = 42.7%, meaning that MobileNetV2 reduces computational resource usage by approximately 42.7% (≈43%) compared to ResNet50.

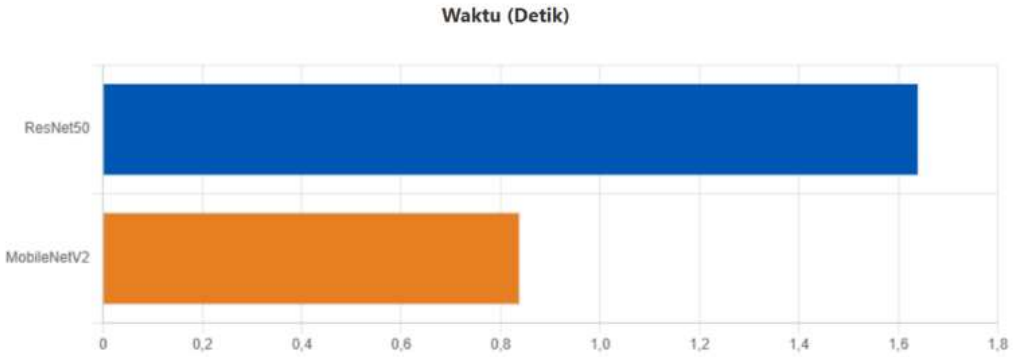

Figure 4. Time Comparison using Augmentation

Figure 4 shows that ResNet50 without Test-Time Augmentation (TTA) requires a feature extraction time of 0.4589 seconds, which increases to 1.6416 seconds when TTA is applied, indicating that the processing time becomes approximately 3.5 times slower. In comparison, MobileNetV2 requires 0.2630 seconds without TTA and 0.8392 seconds with TTA, corresponding to an increase of approximately 3.2 times.

This increase in processing time is both logical and nearly linear. Since TTA expands the inference process to six images (one original image plus five augmented versions), the total extraction time increases by approximately three to four times, accounting for additional processing overhead. Despite this increase, MobileNetV2 with TTA (0.8392 seconds) remains faster than ResNet50 without TTA, highlighting the computational efficiency of MobileNetV2 even under more demanding inference conditions.

Figure 5. Similarity Comparison without Augmentation & threshold 0.5

In the confidence score analysis, as shown in Figure 5, ResNet50 achieves a confidence score of 0.9522, while MobileNetV2 attains a score of 0.9139. ResNet50 demonstrates a very high level of confidence (nearly 97%), indicating that the feature vectors generated by ResNet50 are rich and detailed, enabling the model to distinguish subtle visual similarities with strong discrimination.

In contrast, MobileNetV2 produces a lower maximum confidence score than ResNet50. This outcome is expected, as the compressed architecture of MobileNetV2 results in the loss of some fine-grained visual information, making the model slightly less confident compared to ResNet50. These findings suggest that ResNet50 is more effective in differentiating product variants with highly similar shapes, whereas MobileNetV2 prioritizes computational efficiency. A summary of the comparative results from the single-test evaluation is presented in Table 1.

Table 1. Comparison of Single Test Results

Analysis Parameter	ResNet50 (Non-TTA)	MobileNetV2 (Non-TTA)	ResNet50 (With TTA)	MobileNetV2 (With TTA)	Analysis & Conclusion
Feature Extraction Time	0.4589 seconds (Fairly Fast)	0.2630 seconds (Very Fast)	1.6416 seconds (Slow)	0.8392 seconds (Moderate)	Test-Time Augmentation (TTA) increases computational cost by approximately 3.5×. MobileNetV2 with TTA remains feasible for real-time applications, whereas ResNet50 with TTA becomes computationally heavy (>1.5 s).
Top-1 Retrieval Result	Correct model (NB06179Z – Blue); incorrect color but correct shape	Incorrect model (NB06178Z – Black); correct color only	Incorrect model (NB06181Z – Black); confused by visual twin	Incorrect model (NB06181Z – Black); confused by visual twin	Without TTA, ResNet50 demonstrates superior shape recognition. With TTA, both models converge toward visually similar products (181Z), which have the highest average feature similarity.
Visibility of Correct Product	Very good; appears in Top-1 and Top-3	Poor; does not appear in Top-5	Decreases to Top-5	Improves; appears in Top-3	TTA acts as a double-edged sword. It helps MobileNetV2 recover previously missing correct

Analysis Parameter	ResNet50 (Non-TTA)	MobileNetV2 (Non-TTA)	ResNet50 (With TTA)	MobileNetV2 (With TTA)	Analysis & Conclusion
					products, but degrades ResNet50 by lowering the rank of the correct product.
Visual Feature Focus	Geometry/shape; captures unique frame curvature	Color/texture; mainly matches dominant color	Unique features diluted by averaging	Begins to recognize frame geometry	TTA forces MobileNetV2 to focus more on shape rather than color. Conversely, TTA causes feature dilution in ResNet50, weakening distinctive cues and increasing confusion with distractor products.
Top-1 Similarity Score	0.9522	0.9139	0.9477	0.9283	ResNet50's similarity score slightly decreases with TTA due to less sharp averaged features, while MobileNetV2's score increases as TTA enhances confidence in shape-based recognition.

The batch test was carried out with 11 batch data compared to 4500 train images.

HASIL AKHIR (Top-10)

Query	R-Match	R-Sim	R-Time	R-AP	R-Rec	R-F1	R-RR	M-Match	M-Sim	M-Time	M-AP	M-Rec	M-F1	M-RR
AAGB1-37-90...	AAGB1-37-90...	1.0000	0.9036	0.5382	0.1333	0.2000	0.1111	AAGB1-37-90...	1.0000	0.6347	1.0000	0.1000	0.1500	0.3333
AAGB1-37-90...	AAGB2-37-90...	0.8979	0.1462	0.0000	0.0000	0.0000	0.0000	AAGB2-37-90...	0.8822	0.1009	0.0000	0.0000	0.0000	0.0000
AAGB1-60053...	ANBL1-NB092...	0.9550	0.1425	0.0000	0.0000	0.0000	0.0000	AAGB1-AB600...	0.9270	0.0947	0.0000	0.0000	0.0000	0.0000
AAGB2-AB400...	AAGB1-AB400...	0.9302	0.1162	0.0000	0.0000	0.0000	0.0000	ARYS1-0RX89...	0.9076	0.0680	0.2000	0.0039	0.0075	0.2000
AAGB2-AB701...	AAGB2-AB701...	0.9435	0.1455	1.0000	0.0094	0.0187	0.1000	AAGB2-AB701...	0.9225	0.0989	1.0000	0.0085	0.0168	0.1111
AAGB2-AB701...	AAGB2-AB701...	0.9369	0.1432	0.8412	0.0085	0.0168	0.1000	AAGB2-AB701...	0.9098	0.0983	1.0000	0.0094	0.0187	0.1000
AAGB4-AB100...	AAGB4-AB100...	0.9042	0.1199	0.9283	0.0217	0.0425	0.1000	AAGB4-AB100...	0.8793	0.0671	1.0000	0.0193	0.0377	0.1250
ADLN1-ADLEN...	ADLN1-ADLEN...	0.8355	0.1247	1.0000	0.3333	0.5000	0.1000	ADLN1-ADLEN...	0.8392	0.0800	1.0000	0.3333	0.5000	0.1000
ANBL1-06108...	ANBL16097C0...	0.9312	0.1138	0.3750	0.1667	0.1818	0.1250	ANBL1-NB090...	0.8908	0.0680	0.5000	0.0833	0.0909	0.5000
ANBL1-MB093...	ANBL1-NB093...	0.9295	0.1419	0.0000	0.0000	0.0000	0.0000	ANBL1-NB093...	0.9154	0.0938	0.0000	0.0000	0.0000	0.0000
ANBL1-NB061...	ANBL1-NB061...	0.9453	0.1281	0.8386	0.0111	0.0219	0.1000	ANBL1-NB061...	0.9160	0.0784	1.0000	0.0139	0.0274	0.1000
RATA-RATA		- 0.9281	0.2023	0.5019	0.0622	0.0892	0.0669		- 0.9082	0.1348	0.6091	0.0520	0.0772	0.1427

Figure 6. Batch Test Results

Based on the results illustrated in Figure 6, the evaluation reveals a clear contrast between the deep network architecture (ResNet50) and the lightweight network (MobileNetV2). In terms of computational efficiency, MobileNetV2 demonstrates a significant advantage, achieving an average feature extraction time of 0.1348 seconds, which is approximately 33% faster than ResNet50, whose average extraction time is 0.2023 seconds. A cold-start effect, characterized by a time spike during the first iteration, is observed in both models; however, MobileNetV2 reaches stable inference times much more rapidly in subsequent iterations. This behavior makes MobileNetV2 a more suitable candidate for real-time systems or deployment on resource-constrained devices.

From the perspective of ranking quality, the experimental results present an interesting outcome in which MobileNetV2 outperforms ResNet50. The Mean Average Precision (mAP) achieved by MobileNetV2 is 0.6091, exceeding that of ResNet50, which attains only 0.5019. This superiority is further reinforced by the

Mean Reciprocal Rank (MRR) value of 0.1427 for MobileNetV2, which is nearly twice that of ResNet50. These findings indicate that despite its simpler architecture, MobileNetV2 is more effective in positioning relevant products at higher ranks (Top-1 to Top-3), thereby delivering a more accurate and refined retrieval experience compared to ResNet50, which tends to produce more mixed rankings with less relevant results.

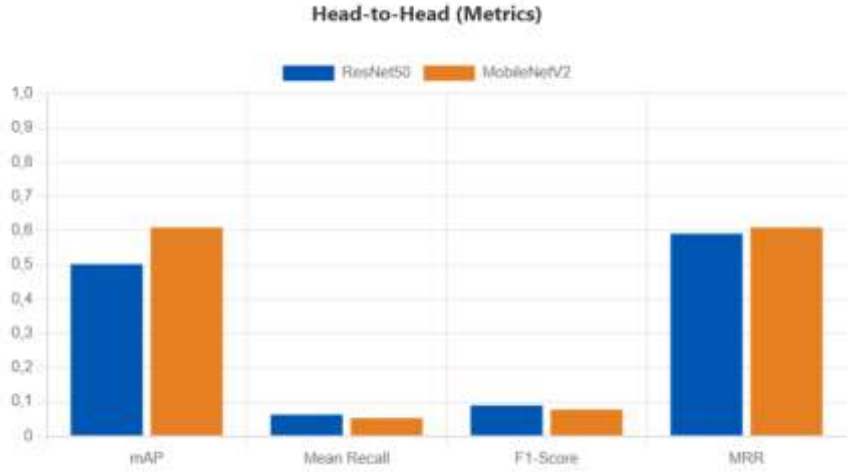

Figure 7. Head-to-Head Analysis Results

On the other hand, as illustrated in Figure 7, ResNet50 demonstrates superiority in terms of search coverage or Recall, achieving an average value of 0.0622 compared to 0.0520 for MobileNetV2. The higher Recall obtained by ResNet50 indicates that its deeper architecture is more capable of capturing complex and abstract visual features, enabling it to retrieve a broader range of relevant product images from the database. However, these relevant images are often distributed across lower ranking positions rather than appearing at the top of the retrieval list.

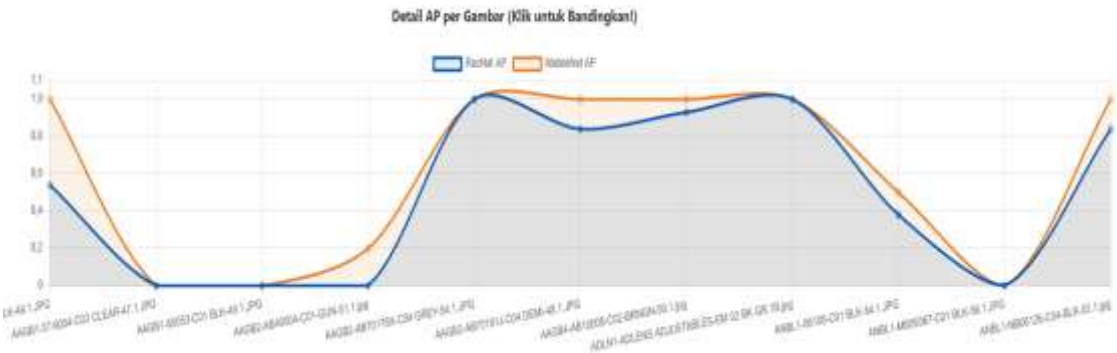

Figure 8. Average Precision Seconds

Figure 8 shows that, in this evaluation sample, MobileNetV2 consistently delivers better or at least comparable ranking quality in terms of Average Precision compared to ResNet50. There are almost no instances where ResNet50 clearly outperforms MobileNetV2, as indicated by the absence of significant gaps where the blue curve rises far above the orange curve. MobileNetV2 demonstrates strong ranking behavior by placing correct retrieval results at the top positions, reflected by Average Precision values approaching 1.0. A summary of the comparative results from the batch testing is presented in Table 2.

Table 2. Comparison of Batch Test Results

Evaluation Parameter	ResNet50	MobileNetV2	Results and Conclusions
Feature Extraction Time	0.2023 seconds	0.1348 seconds	MobileNetV2 shows a significant advantage (~33% faster), making it highly suitable for systems requiring low latency and real-time responses.

Evaluation Parameter	ResNet50	MobileNetV2	Results and Conclusions
Mean Average Precision (mAP)	0.5019	0.6091	MobileNetV2 more consistently ranks relevant products at top positions. The per-image AP curve indicates MobileNetV2 more frequently maintains values close to 1.0 compared to ResNet50.
Mean Reciprocal Rank (MRR – Top-1 Accuracy)	0.0669	0.1427	MobileNetV2 has approximately twice the probability of placing the correct result at the top rank, which is critical for user experience.
Recall	0.0622	0.0520	ResNet50 retrieves a broader range of relevant product variations, but they are often positioned at lower ranks (Top-5 to Top-10).
Similarity Score	0.9281	0.9082	ResNet50 tends to assign high similarity scores even for incorrect matches, while MobileNetV2 is more conservative. This implies that ResNet50 requires a stricter similarity threshold.
Performance Stability	Fluctuating	Stable	Based on the detailed AP curves, ResNet50 frequently exhibits drops in matching regions, indicating intrusion of irrelevant images among correct results. MobileNetV2 demonstrates cleaner and more stable ranking behavior.

In summary, the model usage recommendations can be concluded as follows. MobileNetV2 is more suitable for e-commerce applications or general product catalogs, where users prioritize fast response times and highly relevant results at first glance (Top-3). MobileNetV2 offers the best balance between computational efficiency and clean ranking quality, making it ideal for user-centric retrieval systems. In contrast, ResNet50 is better suited for forensic analysis or inventory auditing applications, where the system must not miss any potential matching products and high recall is the primary requirement. In such contexts, longer processing time and deeper result inspection are acceptable trade-offs. For typical e-commerce scenarios, however, users are unlikely to tolerate slower response times or the need to browse through lower-ranked results, making MobileNetV2 the more practical choice.

5. Conclusion

The experimental results demonstrate that MobileNetV2 consistently delivers superior performance in terms of ranking quality, as reflected by its higher Mean Average Precision (mAP = 0.6091) and Mean Reciprocal Rank (MRR = 0.1427) compared to ResNet50. These findings indicate that MobileNetV2 is more effective at placing relevant products at the top of the retrieval list, which is a critical requirement for user-oriented visual search applications. From a computational efficiency perspective, MobileNetV2 shows a significant advantage, achieving feature extraction times that are approximately 33% faster than those of ResNet50, while also exhibiting more stable inference performance. This efficiency makes MobileNetV2 particularly suitable for real-time systems and environments with limited computational resources. In contrast, ResNet50 outperforms MobileNetV2 in terms of recall, indicating a stronger ability to retrieve a broader range of relevant product candidates from the database. However, this advantage comes at the cost of noisier rankings and a tendency toward overconfidence in similarity scores, which necessitates stricter threshold tuning to avoid irrelevant results. Overall, this study concludes that MobileNetV2 represents a more optimal choice for CBIR systems in eyewear e-commerce and visual catalog applications, where fast response times and accurate top-ranked results are essential. Meanwhile, ResNet50 is better suited for analytical or stock auditing scenarios, where maximizing retrieval coverage is prioritized over

ranking precision. These findings highlight that lightweight CNN architectures can effectively rival deeper models when ranking quality and operational efficiency are the primary objectives.

6. References

[1] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, 2017.

[2] M. W. Sardjono, V. Ramadhan, M. Cahyanti, and E. R. Swedia, "Klasifikasi Bentuk Bingkai (Frame) Kacamata Menggunakan CNN dengan Arsitektur Inception V3 dan Augmented Reality Berbasis Android," *J. Syst. Comput. Eng.*, vol. 5, no. 2, pp. 204–218, 2024.

[3] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[4] T. S. Warongan, S. R. U. A. Sompie, and A. Jacobus, "Penerapan Metode Content-Based Image Retrieval untuk Pengenalan Jenis Bunga," *J. Tek. Inform.*, vol. 13, no. 3, 2018.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] J. Liang, "Image classification based on RESNET," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 12110.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[10] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[11] Y. Gulzar, "Fruit image classification model based on MobileNetV2 with deep transfer learning technique," *Sustainability*, vol. 15, no. 3, p. 1906, 2023.

[12] D. R. Fauzi, "Comparison of CNN Models Using EfficientNetB0, MobileNetV2, and ResNet50 for Traffic Density with Transfer Learning," *J. Intell. Syst. Technol. Informatics*, vol. 1, no. 1, pp. 22–30, 2025.

[13] S. Karnila, S. Irianto, and R. Kurniawan, "Face recognition using content based image retrieval for intelligent security," *Int. J. Adv. Eng. Res. Sci.*, vol. 6, no. 1, pp. 91–98, 2019.

[14] M. R. M. Ariefwan, I. Diyasa, and K. M. Hindrayani, "InceptionV3, ResNet50, ResNet18 and MobileNetV2 performance comparison on face recognition classification," *Literasi Nusant.*, vol. 4, pp. 1–10, 2021.

[15] S. Aras and A. Setyanto, "Deep Learning Untuk Klasifikasi Motif Batik Papua Menggunakan EfficientNet dan Trasnfer Learning," *Insect (Informatics Secur. J. Tek. Inform.*, vol. 8, no. 1, pp. 11–20, 2022.

[16] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[17] F. Loekman, "Sistem Manajemen Inventori Dengan Pengenalan Barang Secara Otomatis Menggunakan Metode Convolutional Neural Network," *Teknika*, vol. 12, no. 1, pp. 47–56, 2023.