




**Paper Type: Original Article**

## Comparative Analysis of Fine-Tuned Pre-Trained Models for Person Re-Identification on the Market-1501 Dataset

**Mohammed A. Fouad <sup>1\*</sup> , Hanaa M. Hamza <sup>1</sup>  and Khalid M. Hosny <sup>1</sup> **

<sup>1</sup> Department of Information Technology, Faculty of Computer and Informatics, Zagazig University, Zagazig 44519, Egypt.  
Emails: [ma.fouad025@fci.zu.edu.eg](mailto:ma.fouad025@fci.zu.edu.eg); [hanaa\\_hamza2000@yahoo.com](mailto:hanaa_hamza2000@yahoo.com); [k\\_hosny@zu.edu.eg](mailto:k_hosny@zu.edu.eg).

**Received:** 25 Aug 2025**Revised:** 04 Nov 2025**Accepted:** 04 Dec 2025**Published:** 11 Dec 2025

### Abstract

Person re-identification (Re-ID) plays a pivotal role in intelligent surveillance, enabling consistent identification of individuals across non-overlapping camera views. Despite the widespread adoption of deep learning, the comparative performance of modern pre-trained architectures under consistent fine-tuning conditions remains underexplored. This study presents a systematic evaluation of five widely used models—ResNet50, DenseNet121, EfficientNetB3, Vision Transformer (ViT), and Swin Transformer—fine-tuned on the Market-1501 dataset using a unified training pipeline. The models were assessed using Rank-1 and Rank-5 accuracy, mean Average Precision (mAP), and computational efficiency metrics such as GFLOPs, FPS, and parameter count. The Swin Transformer achieved the highest Rank-1 accuracy of 96.2% and mAP of 89.1%, outperforming convolutional counterparts while maintaining a competitive inference speed. The results on this benchmark reveal that transformer-based architectures demonstrate superior feature generalization and robustness against viewpoint and illumination variations. The study provides a reproducible benchmark that connects architectural design principles with Re-ID performance, offering practical guidance for future research and deployment in real-time surveillance systems.

**Keywords:** Person Re-Identification; Re-ID; Fine-Tuning; Deep Learning; CNNs; ViT; Pre-Trained Models.

## 1 | Introduction

Person Re-Identification (Re-ID) refers to the task of recognizing and matching individuals across multiple non-overlapping camera views. It has become a core problem in computer vision, with direct applications in intelligent surveillance, urban security, and video analytics. The goal is to accurately associate pedestrian images captured from different angles, backgrounds, and lighting conditions despite significant intra-class variations. However, this task remains highly challenging due to occlusion, pose misalignment, illumination inconsistency, and background clutter, all of which hinder reliable feature extraction and matching performance [1, 2].

In recent years, deep convolutional neural networks (CNNs) have dramatically advanced Re-ID accuracy by learning discriminative appearance representations [3-6]. Architectures such as ResNet and DenseNet improved gradient propagation and feature reuse, while EfficientNet introduced compound scaling to balance accuracy and computational efficiency. More recently, Vision Transformers (ViT) and hierarchical variants such as the Swin Transformer have demonstrated superior global context modeling capabilities and stronger



Corresponding Author: [ma.fouad025@fci.zu.edu.eg](mailto:ma.fouad025@fci.zu.edu.eg)



Licensee International Journal of Computers and Informatics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

robustness to viewpoint variation and background noise [7-10]. Despite these advances, domain shift across cameras continues to degrade generalization performance [11], while attention-based models still struggle under severe occlusion conditions [12]. In parallel, large-scale surveys highlight the rapid expansion of datasets, modalities, and evaluation protocols in Re-ID, indicating a need for more standardized and fair comparison pipelines [13, 14]. Motivated by these gaps, this study conducts a controlled comparison of representative CNN and transformer-based architectures using a unified training and evaluation pipeline.

To this end, we establish a consistent fine-tuning and evaluation pipeline to compare five representative architectures: ResNet50, DenseNet121, EfficientNetB3, Vision Transformer (ViT), and Swin Transformer. The comparative analysis considers Rank-1 and Rank-5 accuracy, mean Average Precision (mAP), and computational efficiency metrics, including GFLOPs, FPS, parameter count, and training duration.

The main contributions of this paper are as follows:

- A unified evaluation framework ensuring fair comparison by maintaining identical data preprocessing, hyperparameters, hardware and a fixed set of standard, literature-based hyperparameters, rather than tuning on a validation set.
- Comprehensive quantitative analysis linking model architecture to performance trade-offs between accuracy and computational cost.

The remainder of this paper is organized as follows. Section 2 reviews related works on person Re-ID and comparative model studies. Section 3 details the datasets, pre-trained models, fine-tuning process, and evaluation metrics. Section 4 presents and discusses the experimental results. Section 5 concludes the paper with key findings and future research directions. All evaluations are conducted using the standard single-query protocol for Market-1501 without any post-processing re-ranking techniques to ensure a fair comparison of the backbone architectures.

## 2 | Related Work

Person Re-Identification (Re-ID) has evolved significantly over the past decade, progressing from hand-crafted feature extraction and metric learning to deep learning-based methods that learn discriminative embeddings end-to-end. Early approaches relied on low-level color histograms, texture descriptors, and distance metrics, which struggled to generalize under varying illumination, pose, and occlusion conditions. The introduction of deep convolutional neural networks (CNNs) revolutionized this field, enabling models to learn hierarchical and invariant representations that improved robustness in camera-to-camera matching [1].

### 2.1 | Deep CNN Architectures for Re-ID

Deep convolutional neural networks have significantly advanced Re-ID performance by learning discriminative embeddings from pedestrian images. ResNet introduced residual skip connections to stabilize gradient flow in deep feature extractors [3], while DenseNet further improved gradient propagation through dense connectivity between layers [4]. EfficientNet later optimized network width, depth, and resolution using compound scaling, improving computational efficiency without sacrificing representation quality [5]. These CNN architectures remain widely used as robust backbones in person Re-ID systems due to their strong balance between accuracy and inference speed.

### 2.2 | Transformer-Based Approaches

Transformer-based architectures have recently reshaped Re-ID research by leveraging global self-attention to capture long-range feature dependencies. Vision Transformers (ViT) apply attention to patch embeddings, enabling holistic spatial reasoning across the full body image [7], while Swin Transformers incorporate hierarchical shifted-window attention to reduce computational overhead and improve scalability [8-10]. Recent efforts have further enhanced transformer-based Re-ID through occlusion-aware modeling [12],

multi-granularity representation learning [15], and large-scale self-supervised training [16]. Survey studies also note that transformers increasingly outperform CNNs on challenging benchmarks, especially when viewpoint variation and background noise are significant [17].

## 2.3 | Comparative and Benchmarking Studies

Although numerous architectures have been proposed, relatively few works have offered fair benchmarking under standardized configurations. Prior evaluations often varied in preprocessing, hyperparameters, or dataset splits, limiting the comparability of results. Recent surveys and benchmarking studies have emphasized the need for unified Re-ID evaluation protocols, deeper analysis of domain shift, and broader consideration of sensor modalities [11, 13, 18-20]. These studies reinforce that architectural improvements must be evaluated alongside real-world deployment factors such as generalization, computation cost, and multi-modal applicability.

## 2.4 | Motivation for This Study

Despite extensive progress, existing Re-ID comparisons often lack uniformity in experimental design, leading to inconsistent conclusions across the literature. Survey analyses further highlight persistent challenges such as domain shift and occlusion and call for systematic comparisons and reproducible training pipelines [11, 13]. Motivated by these findings, this work provides a controlled and transparent benchmark of five pre-trained architectures, linking design principles to accuracy and computational trade-offs.

# 3 | Methodology

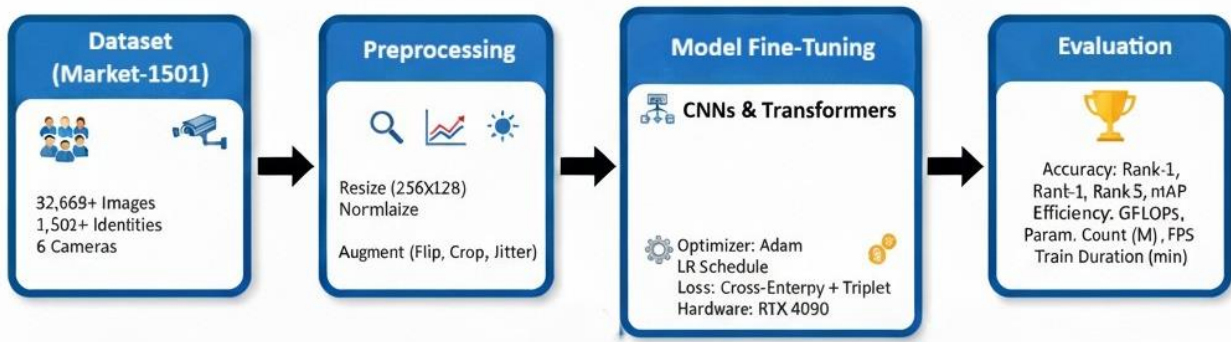
This section outlines the systematic framework for the comparative analysis, as visualized in Figure 1. The framework begins with a description of the Market-1501 dataset and the standardized preprocessing pipeline applied to all images. Subsequently, it introduces the five selected pre-trained architectures and covers their core design principles. The unified fine-tuning procedure, including hyperparameters and loss functions, is then outlined to ensure reproducibility. Finally, the section defines the comprehensive set of accuracy-based and computational efficiency metrics used for the holistic evaluation of the models.

## 3.1 | Datasets

Experiments were conducted on the Market-1501 dataset [8], a widely recognized benchmark for person re-identification research. The dataset consists of 32,668 labeled images representing 1,501 unique identities captured across six cameras with non-overlapping fields of view. It is divided into the standard split of 12,936 training images from 751 identities and a test set (gallery and query) of 19,732 images from the remaining 750 identities. This standard protocol ensures there is no identity overlap between the training and test sets. We use the provided test queries and gallery for evaluation. Figure 2 shows samples of the Market-1501 benchmark dataset.

Images were automatically detected using a Deformable Part Model (DPM) pedestrian detector, resulting in imperfect bounding boxes and background interference — thus providing realistic surveillance conditions. All images were resized to 256×128 pixels and converted to RGB color-space. We applied standard augmentations including random horizontal flipping, random cropping (with padding), and color jittering. Finally, images were normalized using the ImageNet standard mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

This preprocessing ensured a uniform data pipeline across convolutional and transformer-based backbones.



**Figure 1.** The unified experimental pipeline, including preprocessing, model fine-tuning, and evaluation stages.



**Figure 2.** Sample Images from the Market-1501 Dataset.

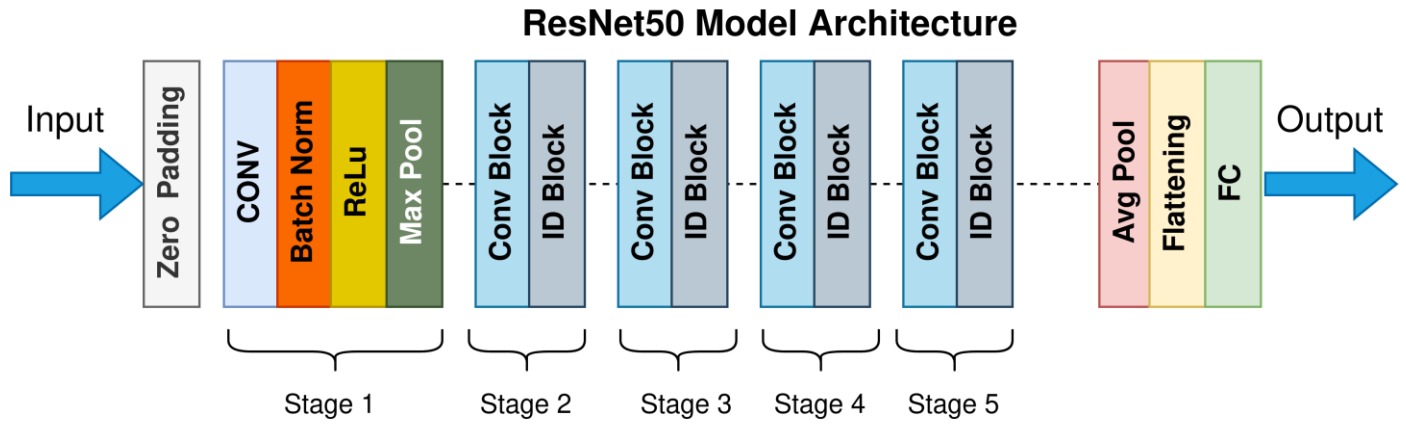
### 3.2 | Pre-Trained Models

Five state-of-the-art architectures were selected for comparative evaluation, representing both convolutional and transformer-based paradigms:

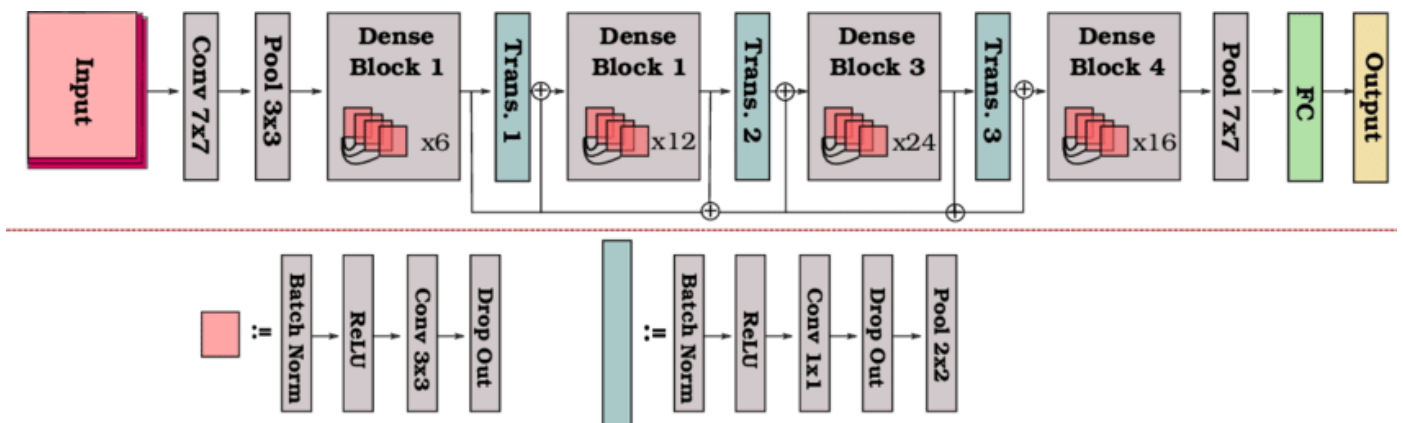
- ResNet50 [3]: A deep residual network introducing skip connections to facilitate gradient flow and stable training of very deep models, as illustrated in Figure 3.
- DenseNet121 [4]: A network employing dense connectivity between layers to maximize feature reuse and efficiency, with its structure shown in Figure 4.
- EfficientNetB3 [5]: A compound-scaled CNN optimized for a balance between accuracy and computational cost, depicted in Figure 5.
- Vision Transformer (ViT-B/16) [6]: A patch-based transformer that applies multi-head self-attention for global context modeling, the framework of which is detailed in Figure 6.
- Swin Transformer (Swin-T) [7]: A hierarchical transformer utilizing shifted windows to capture both local and global information with computational efficiency, as represented in Figure 7.

All models were initialized with ImageNet-pretrained weights to accelerate convergence and improve feature generalization. The architectures were chosen based on their diversity in design philosophy—ranging from purely convolutional to fully attention-based—allowing analysis of how structural differences affect Re-ID performance.

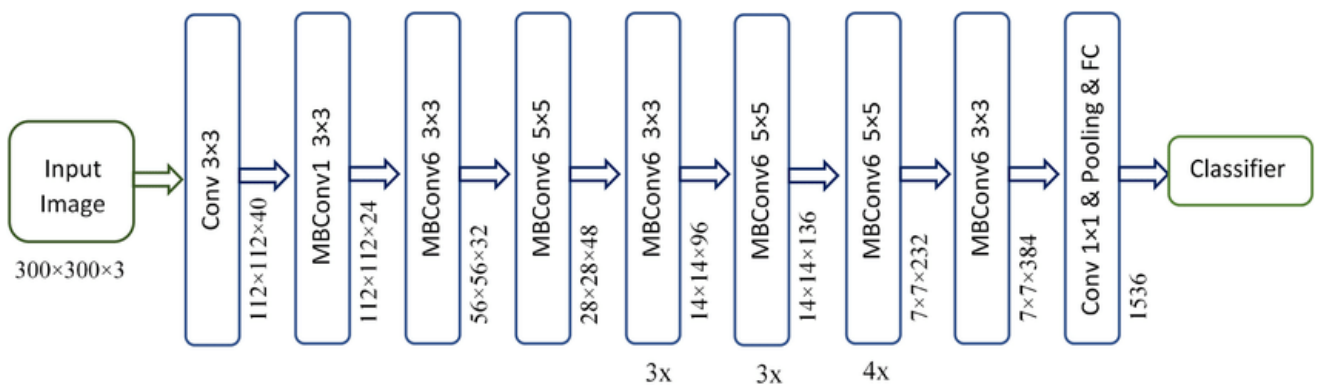
For fine-tuning, the original ImageNet classifier (1000 classes) was replaced. We added a new classification head matching the number of training identities (751 for Market-1501) to be used with the cross-entropy loss. Following standard practice, the final 2048-dimension (for CNNs) or 768-dimension (for Transformers) feature embedding, before this classifier, is used as the identity representation for retrieval during testing.



**Figure 3.** Schematic of Resnet50 architecture showing residual blocks and skip connections enabling deep feature extraction [3].

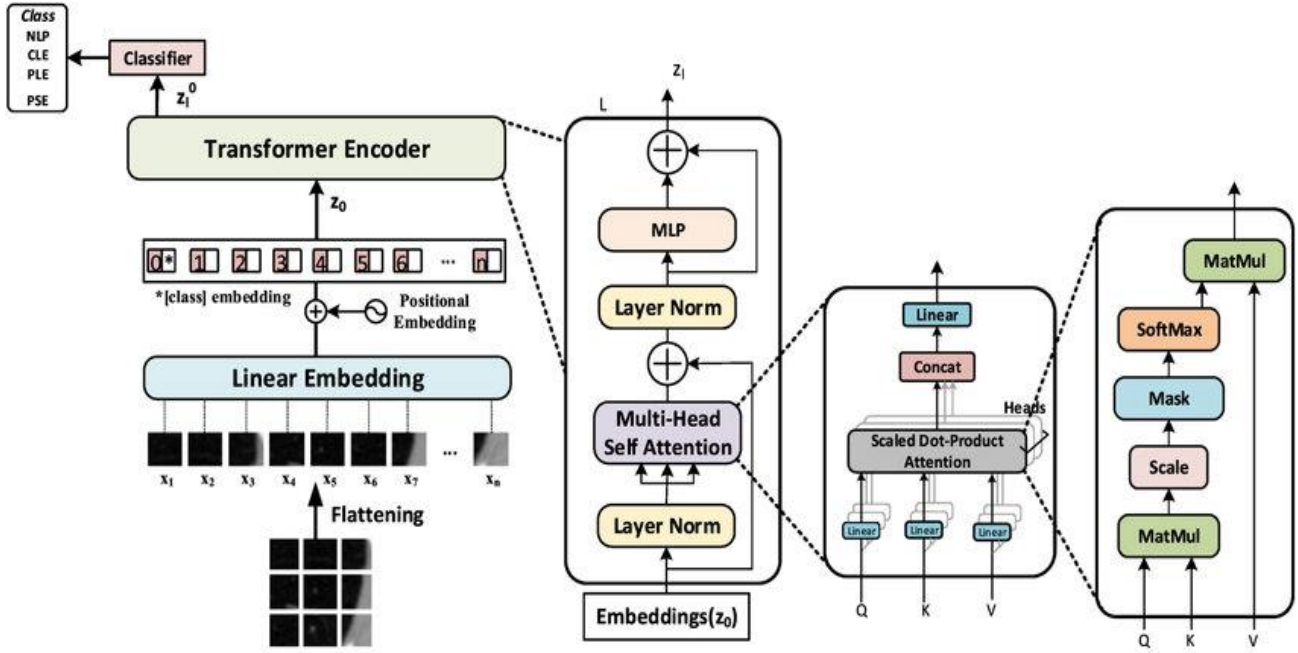


**Figure 4.** Schematic of Densenet121 highlighting dense connectivity that concatenates feature maps from preceding layers [4].

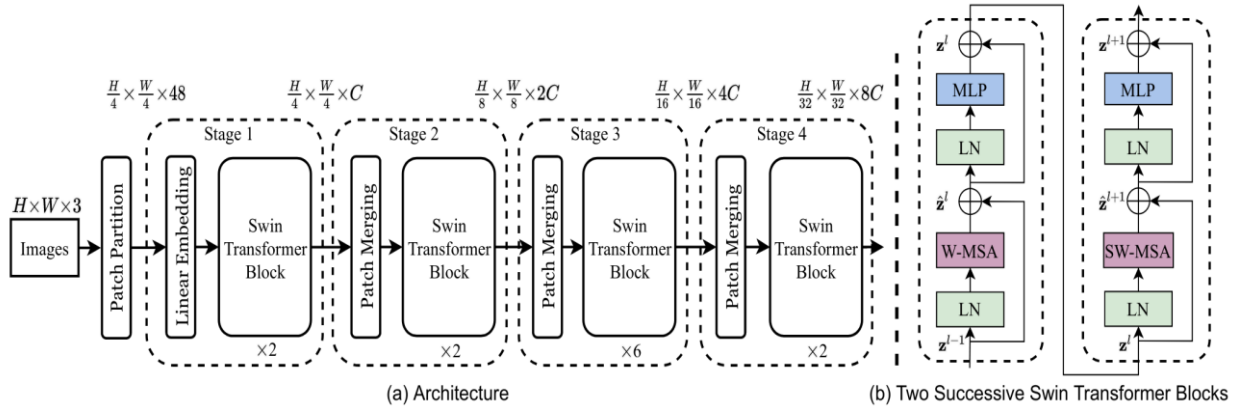


**Figure 5.** Structural overview of EfficientNetB3 depicting compound scaling of depth, width, and resolution [5].





**Figure 6.** Vision Transformer (ViT-B/16) framework illustrating image patch embedding and multi-head self-attention mechanism [6].



**Figure 7.** Swin Transformer (Swin-T) representation demonstrating hierarchical shifted-window attention for efficient global context modeling [7].

### 3.3 | Fine-Tuning Procedure

All models were fine-tuned using a unified pipeline to ensure reproducibility and fairness. Training employed the Adam optimizer with a weight decay of  $5 \times 10^{-4}$  and an initial learning rate of  $3 \times 10^{-4}$ , reduced by a fixed schedule (a factor of 0.1 every 20 epochs) to avoid any adaptive changes based on test-set feedback. All models were trained on the full Market-1501 training set for a fixed 120 epochs with a batch size of 32.

The objective combined cross-entropy loss for identity classification with a hard-mining Triplet Loss (with a margin  $m=0.3$ ) to enhance inter-class separation. The triplet loss was computed using 'batch hard' mining, where for each anchor in the batch, the hardest positive and hardest negative samples within that batch are selected.

In the early epochs, lower convolutional or transformer layers were frozen to retain general visual features, and unfreezing was gradually introduced to enable full network optimization. All experiments were executed on a P100 GPU (16 GB VRAM) with a fixed random seed (seed=42) for PyTorch, CUDA, and NumPy, and enabled cuDNN deterministic flags to ensure consistent results across runs.

### 3.4 | Evaluation metrics

A holistic evaluation framework was adopted to comprehensively assess the recognition performance and computational efficiency of all tested models. The framework integrates accuracy-oriented retrieval metrics that assess the correctness of identity matching with computational-efficiency indicators that quantify model complexity and runtime performance. Together, these metrics ensure a balanced assessment of both predictive power and resource utilization. The adopted evaluation criteria are summarized in Table 1.

**Table 1.** Evaluation metrics used for model assessment.

Category	Metric	Description
Accuracy-based	Rank-1 Accuracy	Percentage of queries where the correct identity ranks first in the retrieval list.
	Rank-5 Accuracy	Probability that the correct identity appears within the top five ranked results.
	Mean Average Precision (mAP)	Overall retrieval performance is computed as the mean of the average precision across all queries.
Computational Efficiency	GFLOPs	Number of floating-point operations per forward pass, representing computational complexity.
	FPS	Frames processed per second during inference, reflecting runtime efficiency.
	Parameter Count (M)	Total number of trainable parameters (in millions), indicating model capacity and storage demand.
	Training Duration (min)	Time required for full convergence under identical hardware and configurations.
Protocol Settings	All accuracy metrics are calculated using the standard single-query protocol without re-ranking. mAP calculation excludes 'junk' images.	

#### 3.4.1 | Accuracy-Oriented Metrics

##### (a) Rank-1 and Rank-5 Accuracy:

These top- $k$  retrieval measures evaluate how frequently the correct identity appears among the top predictions within the gallery set. For a collection of query images  $Q = \{q_1, q_2, \dots, q_N\}$ , the Rank- $k$  accuracy is mathematically expressed as:

$$\text{Rank-}k = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{GT}(q_i) \in \text{Top-}k(q_i)) \quad (1)$$

where  $\mathbb{I}(\cdot)$  denotes the indicator function,  $\text{GT}(q_i)$  represents the ground-truth identity of the query  $q_i$ , and  $\text{Top-}k(q_i)$  refers to the top- $k$  retrieved results. Rank-1 accuracy (where  $k = 1$ ) indicates exact identification capability, whereas Rank-5 accuracy evaluates near-correct retrieval consistency, both serving as standard benchmarks in Re-ID studies.

##### (b) Mean Average Precision (mAP):

The mAP metric provides a more comprehensive measure of retrieval effectiveness by averaging precision across recall levels for all queries. For each query  $q_i$ , the Average Precision (AP) is defined as:

$$\text{AP}(q_i) = \frac{1}{m_i} \sum_{k=1}^n P(k) \cdot \Delta R(k) \quad (2)$$

where  $P(k)$  denotes the precision at rank  $k$ ,  $\Delta R(k)$  represents the incremental recall change, and  $m_i$  is the number of relevant gallery images for the query  $q_i$ . The mean over all queries yields mAP:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}(q_i) \quad (3)$$

A higher mAP value indicates stronger ranking consistency and robustness, especially when multiple instances of the same identity exist in the gallery. Our evaluation follows the standard Market-1501 protocol, which excludes 'junk' images (distractors or false detections) from the precision calculation, preventing them from penalizing the score.

### 3.4.2 | Computational Efficiency Metrics

**(a) GFLOPs (Giga Floating-Point Operations):** GFLOPs quantify the number of floating-point operations required for one forward pass. This hardware-independent measure reflects the model's algorithmic complexity. GFLOPs were calculated for a single forward pass using a standard input size of  $256 \times 128 \times 3$  with the thop (PyTorch-OpCounter) library, which accounts for both convolutional and attention operations. Lower GFLOPs imply lower computational demand and faster inference, which is crucial for real-time applications.

**(b) Frames Per Second (FPS):** FPS measures inference throughput, representing how many image frames can be processed per second during testing. FPS was measured on the same P100 GPU using FP32 precision. We measured the inference time for 1000 images with a batch size of 1 (to simulate a real-world single-query scenario) *after* an initial warm-up period of 100 iterations to exclude initialization overhead. Data-loader time was not included, to purely measure model inference speed. High FPS values denote better runtime efficiency and suitability for live surveillance systems, where latency is a key performance factor.

**(c) Parameter Count (M):** This metric captures the total number of trainable parameters in millions, offering insight into model capacity and storage requirements. A lower parameter count generally implies a lighter model, though excessive reduction may limit learning capacity and representational power.

**(d) Training Duration:** Training duration measures the total wall-clock time, in minutes, required to complete the full 120-epoch training schedule on the Market-1501 dataset. This measurement was taken on a single P100 GPU and includes all overhead such as data augmentation, I/O, and model backpropagation, providing a practical measure of training cost.

Overall, these metrics jointly provide a comprehensive assessment of model performance, covering not only retrieval accuracy but also computational feasibility. This dual-perspective evaluation ensures that the comparative analysis remains relevant to both research and real-world deployment contexts, where accuracy, speed, and efficiency must coexist.

## 4 | Experimental Results and Discussion

This section presents the quantitative results of the five evaluated models and discusses their comparative performance on the Market-1501 dataset. All models were fine-tuned under identical hyperparameters, data preprocessing, and computational environments to ensure experimental fairness. Crucially, all results are reported without any post-processing re-ranking techniques (e.g., k-reciprocal). This ensures the comparison reflects the raw feature extraction power of the backbones themselves.

### 4.1 | Quantitative Comparison

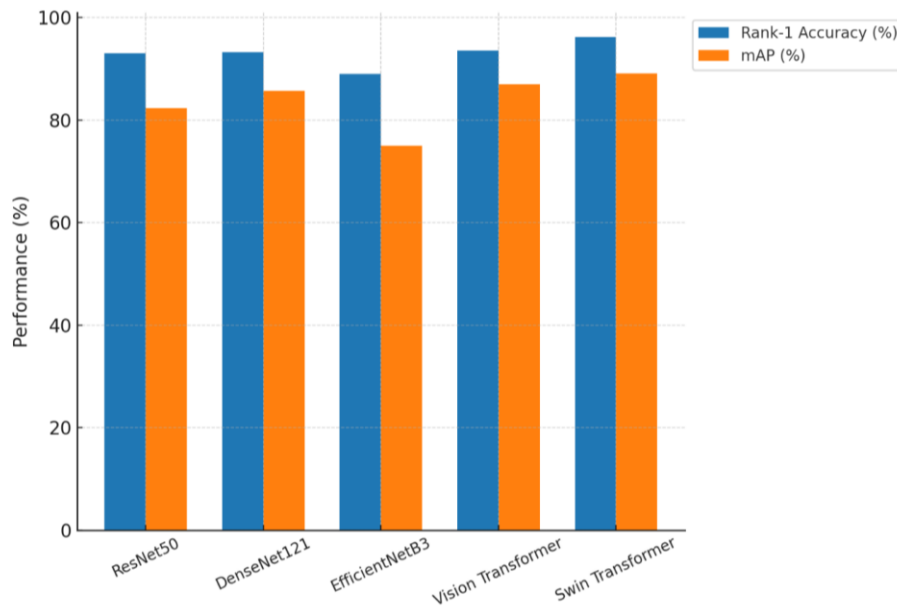
Table 2 summarizes the evaluation results for all architectures, including Rank-1 and Rank-5 accuracies, mean Average Precision (mAP), computational complexity (GFLOPs), inference speed (FPS), parameter count, and training duration.



**Table 2.** Quantitative results of all models fine-tuned on Market-1501.

Model	Rank-1 Accuracy	Rank-5 Accuracy	mAP	GFLOPs	FPS	Parameters (M)	Training Duration (min)
ResNet50	93.0	97.8	82.3	4.1	35	25.6	210
DenseNet121	93.2	98.0	85.7	7.9	25	8.0	230
EfficientNetB3	89.0	97.5	75.0	1.8	45	12.0	180
Vision Transformer	93.5	98.5	87.0	17.6	20	87.0	350
Swin Transformer	96.2	99.0	89.1	15.4	25	65.0	385

The results demonstrate that the Swin Transformer achieves the highest Rank-1 accuracy (96.2%) and mAP (89.1%), outperforming all convolutional counterparts. The Vision Transformer ranks second, slightly surpassing ResNet50 and DenseNet121 in retrieval accuracy. Notably, EfficientNetB3, despite being the most lightweight model with the fewest GFLOPs (1.8) and the highest inference speed (45 FPS), exhibits a lower mAP of 75.0%, highlighting a trade-off between speed and representational power. Among the CNN-based architectures, DenseNet121 marginally outperforms ResNet50 in both Rank-1 accuracy and mAP (93.2% vs. 93.0% and 85.7% vs. 82.3%, respectively) due to its dense connectivity, which maximizes feature reuse and enhances gradient propagation. However, its increased GFLOPs result in slower inference. Transformer-based models (ViT and Swin) require considerably more computational resources and longer training times but yield stronger generalization and better feature separability across varying viewpoints and lighting conditions.

**Figure 8.** Comparison of Rank-1 accuracy and mAP across all evaluated models on the Market-1501 dataset.

## 4.2 | Discussion

The observed performance differences can be attributed to the fundamental architectural principles underlying each model. ResNet50, characterized by residual skip connections, maintains strong baseline performance by facilitating stable gradient flow in deep networks. However, its convolutional receptive fields limit its capacity for modeling global context, which is crucial in Re-ID tasks where fine-grained spatial dependencies span across body parts.

DenseNet121 improves upon this by concatenating feature maps from preceding layers, effectively combining low-level and high-level information. This dense connectivity enhances the discrimination of subtle clothing and texture cues, thereby improving mAP. Nonetheless, DenseNet's internal redundancy slightly increases computational cost and reduces inference speed compared to ResNet50.

EfficientNetB3, built around compound scaling, is optimized for efficiency. Its lower GFLOPs and parameter count make it ideal for real-time applications on edge devices. However, its reduced depth and feature dimension limit its ability to capture complex spatial variations, explaining the noticeable drop in Rank-1 and mAP performance. In contrast, Vision Transformer (ViT) leverages global self-attention to establish long-range dependencies, enabling the model to understand spatial relationships between distant image regions. This results in improved feature representation and robust matching under illumination and pose variations. Yet, ViT's high computational cost (17.6 GFLOPs) and large parameter size (87M) render it less suitable for deployment in latency-sensitive systems.

Finally, the Swin Transformer achieves the best overall balance between accuracy and efficiency. Its hierarchical design and shifted window mechanism combine the benefits of local feature extraction with scalable global context modeling. This allows Swin-T to outperform both ViT and all CNN baselines in Rank-1 and mAP while maintaining moderate GFLOPs and manageable inference speed (25 FPS). The model's longer training duration (385 minutes) reflects the computational cost of hierarchical attention, but the accuracy gains justify its usage for high-stakes surveillance applications. Overall, these results suggest that transformer-based architectures, particularly hierarchical ones like Swin-T, provide superior generalization and robustness for person Re-ID tasks. However, for resource-limited environments, CNNs such as ResNet50 or EfficientNetB3 still offer compelling trade-offs between accuracy and speed. We acknowledge, however, a key limitation of this study. The results presented in Table 2 are based on a single, controlled run (with fixed seeds) for each model due to computational constraints. While this provides a direct comparison, it does not capture the variance across multiple training runs. Future work should include repeated experiments to establish confidence intervals and further strengthen these conclusions.

## 5 | Conclusion

In this study, we presented a controlled and systematic comparison of five pre-trained deep learning architectures, ResNet50, DenseNet121, EfficientNetB3, ViT, and Swin Transformer, fine-tuned for Person Re-Identification on the Market-1501 dataset under identical experimental conditions. Our results provide a clear benchmark for assessing architectural performance trade-offs in Re-ID.

The experimental findings demonstrate a clear advantage for the Swin Transformer, which achieved the highest retrieval performance with a Rank-1 accuracy of 96.2% and an mAP of 89.1%. This confirms that hierarchical transformer architectures excel at capturing the necessary global and local features for robust identity matching across non-overlapping camera views. Conversely, while transformer-based models provided superior accuracy, EfficientNetB3 demonstrated the best computational efficiency, achieving the highest FPS with the lowest GFLOPs. This highlights the critical trade-off: deep attention-based models prioritize representational power over real-time constraints, while compound-scaled CNNs offer a compelling alternative for resource-limited, latency-sensitive deployments.

For future research, we suggest two primary directions:

1. **Efficiency Optimization:** Further investigation into pruning or quantization techniques for the Swin Transformer to reduce its computational footprint while preserving its high retrieval accuracy, making it viable for edge computing devices.
2. **Hybrid Architectures:** Exploration of hybrid models that leverage the localized feature extraction power of EfficientNet with the global context modeling of Swin Transformer blocks, aiming to achieve the 'best of both worlds' in terms of high accuracy and high inference speed.

3. **Cross-Dataset Validation:** Validating this framework on other large-scale Re-ID benchmarks (e.g., DukeMTMC-reID, MSMT17) to test the generalization of these findings, which are currently limited to the Market-1501 dataset.
4. **Qualitative Analysis:** A deeper qualitative analysis of failure cases. This would involve visualizing and analyzing the per-query AP distribution to identify specific challenges (e.g., severe occlusion, rare viewpoints) where models still fail, providing more granular insight beyond mean performance.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Data Availability

The Market-1501 dataset used in this study is publicly available. To ensure full reproducibility, the code, training configurations, and final model checkpoints for all five architectures will be made available upon publication in a public GitHub repository.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

- [1] D. Avola, E. Emam, D. Montagnini, D. Pannone, and A. Ranaldi, "WhoFi: Deep Person Re-Identification via Wi-Fi Channel Signal Encoding," arXiv:2507.12869, 2025.
- [2] M. Song, "Exploring the Camera Bias of Person Re-Identification," OpenReview, 2025.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
- [4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.
- [5] M. T. Ramakrishna, K. Pothanaicker, P. Selvaraj, S. B. Khan, V. K. Venkatesan, S. Alzahrani, and M. Alojail, "Leveraging EfficientNetB3 in a Deep Learning Framework for High-Accuracy MRI Tumor Classification," Computers, Materials & Continua (CMC), vol. 81, no. 1, pp. 867–881, 2024.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021. <https://arxiv.org/abs/2010.11929>
- [7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021. <https://doi.org/10.1109/ICCV48922.2021.00985>
- [8] Liu et al., "Domain Generalization for Person Re-Identification," Neurocomputing, 2025. <https://www.sciencedirect.com/science/article/abs/pii/S0925231225014353>
- [9] T.-D. Nguyen et al., "Tackling Domain Shifts in Person Re-Identification: A Survey and Analysis," in CVPR Workshops, 2024.
- [10] Z. Ji, D. Cheng, and K. Feng, "Exploring Stronger Transformer Representation Learning for Occluded Person Re-Identification," arXiv:2410.15613, 2024.
- [11] X. Wang et al., "A survey on person and vehicle re-identification," IET Computer Vision, 2024.
- [12] A. Wang and Y. Zhang, "Comprehensive Survey on Person Identification: Queries, Datasets, Metrics, and Approaches," Communications of the ACM, vol. 67, no. 6, pp. 32–45, 2024.
- [13] L. Zhang, J. Liu, and Y. Cao, "Learning multi-granularity representation with transformer for person re-identification," Pattern Recognition, vol. 153, p. 110264, 2025.
- [14] Y. Shen et al., "PersonViT: Large-scale Self-supervised Vision Transformer for Person Re-Identification," arXiv:2408.05398, 2024.
- [15] H. Rao and C. Miao, "Recognizing Identities From Human Skeletons: A Survey on 3D Skeleton Based Person Re-Identification," arXiv:2401.15296, 2025.

- [16] K. Wu, X. Li, and L. Ma, "UAV-based person re-identification: A survey of UAV datasets, challenges, and methods," *Computer Vision and Image Understanding (CVIU)*, vol. 233, p. 104575, 2025.
- [17] R. Ha, S. Jiang, B. Li, B. Pan, Y. Zhu, J. Zhang, X. Zhu, S. Gong, and J. Wang, "Multi-modal Multi-platform Person Re-Identification: Benchmark and Method," *arXiv:2503.17096*, 2025.
- [18] X. Wang and Y. Zhang, "A survey on person and vehicle re-identification," *IET Computer Vision*, 2024.
- [19] A. Wang and Y. Zhang, "Comprehensive Survey on Person Identification," *CACM*, 2024.
- [20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-Identification: A Benchmark," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.