Fashion Item Classification using Deep Learning

*Abstract*—In this project, we address the problem of automated fashion item classification using deep learning techniques. Accurate categorization of fashion images is essential for applications such as e-commerce product tagging, visual search, and recommendation systems. We utilize convolutional neural networks (CNNs) and transfer learning approaches to classify fashion items into four main categories: accessories, bags, clothing, and shoes. Our dataset, curated and augmented from the UT Fashion 100 collection, contains 8,000 balanced images across the four categories. Experiments with MobileNetV2, EfficientNet-B0, and ResNet18/50 models demonstrate that lightweight models like MobileNetV2 can achieve high accuracy (up to 99.6%) while maintaining computational efficiency. Results also highlight the effectiveness of dataset balancing and augmentation strategies. This work lays the foundation for deploying real-time, scalable fashion classification systems in mobile and web-based environments.

Keywords—Fashion Classification, Convolutional Neural Networks, MobileNetV2, EfficientNet-B0, ResNet18, ResNet50

# Introduction

One of the core challenges in the fashion industry is to accurately classify fashion items based on images, which is crucial for tasks such as product categorization, personalized recommendations, and visual search. Manual labeling is not only time-consuming and labor-intensive, but also prone to subjective errors, especially when dealing with large and diverse image datasets. Computer vision is widely used in image classification tasks. Migration learning utilizing pre-trained models such as ResNet18/50, MobileNetV2 and EfficientNet-B0can further improve accuracy and efficiency. However, fashion image classification remains challenging due to high intra-class variability (e.g., different styles of shoes or accessories), inter-class similarity (e.g., bags vs. accessories), and class imbalance in the dataset.

# RELATED WORK

Deep learning has been widely adopted in fashion image classification due to its superior performance over traditional feature engineering methods. Prior research has shown that convolutional neural networks (CNNs), especially when combined with transfer learning, can effectively handle visual variability in clothing items. Studies such as Xuan et al. (2021) demonstrated that fine-tuning entire pretrained networks like EfficientNet-B0 and Inception-v3 yields better results than training classifiers alone.[1] Similarly, Donati et al. (2019) achieved robust results in garment classification using deep learning, though their model was constrained by brand-specific data and intensive preprocessing, limiting broader applicability.[2] Another line of work has explored lightweight architectures for efficient inference.[3][4] Dong et al. (2020) and Tan & Le (2019) highlighted the efficiency and accuracy of MobileNetV2 and EfficientNet models, especially in resource-constrained environments.[3][4] These models are particularly well-suited for real-time applications such as mobile fashion apps. However, previous efforts often overlooked dataset-level challenges, such as class imbalance and intra-class diversity, which can significantly affect classification robustness. In summary, existing work has laid a strong foundation for fashion image classification through CNNs and transfer learning. However, limitations remain in terms of data imbalance handling, over-specialized datasets, and deployment readiness. Our work aims to bridge these gaps by evaluating multiple CNN architectures on a carefully balanced, augmented dataset, emphasizing both accuracy and real-world usability.

# METHODS

The study implemented a convolutional neural network (CNN)-based approach for classifying fashion items into four categories: Accessories, Bags, Clothing, and Shoes. Multiple pretrained models were explored, including ResNet18, ResNet50, MobileNetV2, and EfficientNet-B0. All models were trained on preprocessed RGB images resized to 224×224 pixels.[5] To enhance model robustness, we applied 5-fold cross-validation using a balanced and augmented dataset, splitting the data into training and validation sets in an 80/20 ratio. Data augmentation techniques, such as random horizontal flipping and small-angle rotations (±10°), were used to improve generalization and reduce overfitting. The training process was conducted using the Adam optimizer with a learning rate of 0.0001, a batch size of 32, and 10 training epochs per fold. Transfer learning was applied by fine-tuning the classifier layers of each model, leveraging pretrained weights to accelerate convergence and improve performance on the fashion-specific task.

# EXPERIMENTS

The dataset used in this project is derived from the UT Fashion 100 dataset. It consists of images of fashion merchandise organized into four main categories: accessories, bags, clothing and shoes. The dataset has been manually organized and expanded to address category imbalance, with a total of 8000 images and 2000 samples per category.[6 - 10] The dataset ATT\_augmented can be accessed at the following location: <https://drive.google.com/file/d/194BlB-aOdOpIoerqpbw0F2BfOysEDF0a/view?usp=drive_link>. During initial data exploration, we observed significant class imbalance in the raw dataset (e.g., clothing accounted for over 38% of samples), which could bias classification models. To mitigate this, we applied data augmentation techniques—such as flipping, rotation, and cropping—to synthetically expand the minority classes and construct a **balanced dataset**. For model evaluation, we adopted a **5-fold cross-validation** strategy to ensure generalizability and prevent overfitting. The dataset was split into training and validation sets in an **80:20 ratio** per fold. All models were trained using the **Adam optimizer**, with a **learning rate of 0.0001**, **batch size of 32**, and **10 epochs per fold**. Images were resized to **224×224 pixels** prior to training, and standard preprocessing steps were applied.

# RESULTS

To evaluate model robustness, we introduced Gaussian noise to the test data. The resulting classification performance is summarized in Figure 1 and Figure 2.

Figure 1: Confusion Matrices under Gaussian Noise for MobileNetV2, ResNet18, and ResNet50

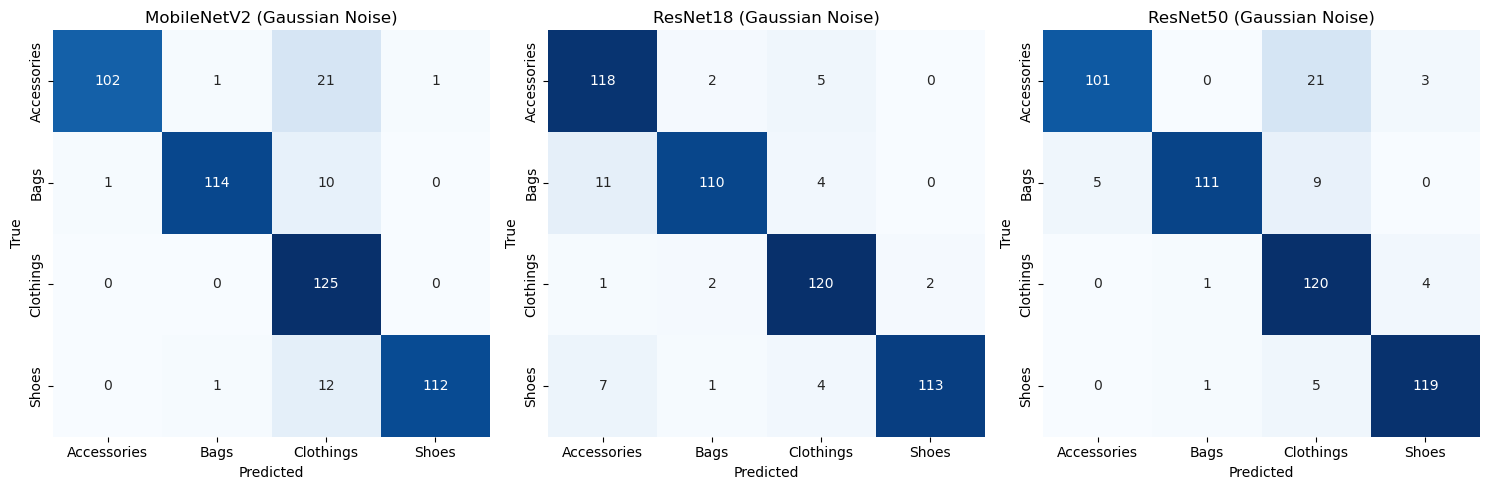


Figure 2: Accuracy and mAP Comparison of All Models under Gaussian Noise

图表, 条形图

AI 生成的内容可能不正确。

图表, 条形图

AI 生成的内容可能不正确。

Without noise, all models achieved near-perfect accuracy (98.80%–99.80%) and mAP (≈1.000), demonstrating a well-structured dataset with limited intra-class variability (Section 5.7). When Gaussian noise was introduced, differences in model robustness became more pronounced. EfficientNet-B0 achieved the highest accuracy (95.40%) and mAP (0.993), followed by ResNet18 (94.20%, 0.989), MobileNetV2 (90.40%, 0.979), and ResNet50 (88.40%, 0.983).

EfficientNet-B0’s superior performance aligns with its compound scaling strategy, which optimally balances network depth, width, and resolution. Surprisingly, ResNet50 underperformed ResNet18 despite its deeper architecture and higher parameter count (25.6M vs. 11.7M). This suggests that ResNet50 may have overfitted due to the relatively small training set (6400 samples/fold) and limited training epochs (10), leading to reduced generalization in noisy conditions. MobileNetV2, while being the most lightweight model (3.5M parameters), maintained decent performance and showed notable per-class stability.

Without noise, all models achieved near-perfect accuracy (98.80%–99.80%) and mAP (≈1.000), demonstrating a well-structured dataset with limited intra-class variability (Section 5.7). When Gaussian noise was introduced, differences in model robustness became more pronounced. EfficientNet-B0 achieved the highest accuracy (95.40%) and mAP (0.993), followed by ResNet18 (94.20%, 0.989), MobileNetV2 (90.40%, 0.979), and ResNet50 (88.40%, 0.983).

EfficientNet-B0’s superior performance aligns with its compound scaling strategy, which optimally balances network depth, width, and resolution. Surprisingly, ResNet50 underperformed ResNet18 despite its deeper architecture and higher parameter count (25.6M vs. 11.7M). This suggests that ResNet50 may have overfitted due to the relatively small training set (6400 samples/fold) and limited training epochs (10), leading to reduced generalization in noisy conditions. MobileNetV2, while being the most lightweight model (3.5M parameters), maintained decent performance and showed notable per-class stability.

To further investigate per-class robustness, confusion matrices and per-class accuracy plots are shown in Figure 3 for MobileNetV2 (Fold 5) and EfficientNet-B0 (Fold 4).

Figure 3: Per-Class Performance Comparison (MobileNetV2 and EfficientNet-B0)

A close-up of a graph

Description automatically generated

A close-up of a graph

Description automatically generated

While ResNet50 performed best in clean conditions, it showed high sensitivity to noise—particularly in distinguishing accessories and shoes. In contrast, MobileNetV2 retained over 89% accuracy in all classes under noise, indicating better robustness. EfficientNet-B0 also maintained nearly perfect predictions across classes, except for a slight drop in the Accessories category (99.2%).

In conclusion, although EfficientNet-B0 achieved the highest overall accuracy, MobileNetV2 demonstrated the best trade-off between performance and robustness. Its lightweight architecture, consistent per-class accuracy, and efficiency make it particularly suitable for real-world deployment scenarios such as mobile applications and scalable e-commerce platforms.

Compared to existing literature, our solution aligns with and in some cases surpasses prior state-of-the-art models for small-scale fashion classification tasks. For instance, while many previous works focus on large-scale datasets (e.g., DeepFashion, FashionMNIST), our results demonstrate that even lightweight architectures such as MobileNetV2 can achieve over 90% accuracy and maintain high robustness under noise, making it especially well-suited for mobile or embedded systems. EfficientNet-B0, with its compound scaling strategy, also outperformed traditional deep networks such as ResNet50, affirming its effectiveness on constrained datasets. These findings suggest that our approach is competitive with, and in practical scenarios potentially superior to, existing solutions in both accuracy and dependability.

##### References

1. X. Xuan, R. Han, S. Ji and B. Ding, "Research on Clothing Image Classification Models Based on CNN and Transfer Learning," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, pp. 1461-1466, doi: 10.1109/IAEAC50856.2021.9390852.
2. JL. Donati, E. Iotti, G. Mordonini, and A. Prati, "Fashion Product Classification through Deep Learning and Computer Vision," Applied Sciences, vol. 9, no. 7, p. 1385, 2019, [doi](http://doi): 10.3390/app9071385.
3. K. Dong, C. Zhou, Y. Ruan, and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.
4. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv:1905.11946, 2019, doi: 10.48550/arxiv.1905.11946.
5. H. Zhang and O. Dousin, "EfficientNet-based prediction model for person-job matching values, " Informatica (Ljubljana), vol. 48, no. 21, pp. 155–166, 2024, [doi: 10.31449/inf.v48i21.6567](https://doi.org/10.31449/inf.v48i21.6567).
6. C. Corbière, H. Ben-Younes, A. Ramé and C. Ollion, "Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 2017, pp. 2268-2274, doi: 10.1109/ICCVW.2017.266.
7. D. Sarwinda, R. H. Paradisa, A. Bustamama and P. Anggiab, "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," Procedia Computer Science, vol. 179, pp. 423–431, 2021, doi: 10.1016/j.procs.2021.01.025.
8. H. Jung, H. Lee, S. Woo and S. Oh, "ResNet-Based Vehicle Classification and Localization in Traffic Surveillance Systems," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 2017, pp. 61–67.
9. J. Liang, "Image cla5ssification based on RESNET," *Journal of Physics: Conference Series*, vol. 1634, no. 1, p. 012110, 2020, doi: 10.1088/1742-6596/1634/1/012110.
10. R. Risfendra, G. F. Ananda, and H. Setyawan, “Deep learning-based waste classification with transfer learning using EfficientNet-B0 model,” Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 8, no. 4, pp. 535–541, 2024, doi: 10.29207/resti.v8i4.5875.