International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Building framework recommendation system for trendy fashion e-commerce based on deep learning with Top-K

Ho Thanh Thuy [1], Pham The Bao [2,*] and Do Dieu Le [1]

[1] *Computer Science Department, Mathematics and Computer Science Faculty, University of Science University, Hochiminh city, Vietnam.*
[2] *Computer Science Department, Information Science Faculty, Sai Gon University, Hochiminh city, Vietnam.*

International Journal of Science and Research Archive, 2024, 12(02), 664–675

Publication history: Received on 02 June 2024; revised on 10 July 2024; accepted on 13 July 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.2.1270

Abstract

Recently, e-commerce has become a vital component of our purchasing habits. Central to this evolution is the recommendation system, an advanced algorithm designed to personalize the shopping experience and significantly boost consumer demand. With its diverse and ever-changing inventory, the fashion industry benefits immensely from these algorithms, making it a fascinating case study for understanding the broader impacts of technology on consumerism. Traditional fashion recommendation systems are fundamentally based on item compatibility, but keeping up with trends is also essential. To address this, we propose a two-stage system: fashion detection and outfit suggestions based on the identified items. Users receive images of Key Opinion Leaders (KOLs) or Influencers wearing similar outfits. These recommendations ensure item compatibility, offer diverse styles, and remain fashionable. At the outset, we experimented with YOLOv8 to select the best version. Next, we implemented fashion image retrieval based on feature extraction using two pre-trained networks. To enhance reliability, we developed a voting and ranking algorithm. Our experiments, conducted on a self-collected dataset, evaluated the system's effectiveness in detecting fashion objects and the efficiency of content-based image retrieval

Keywords: Recommendations system; Deep Learning; Fashion Apparel Detection; YOLOv8; Fashion Image Retrieval

1. Introduction

As users increasingly have easier access to a diverse range of fashion items, it raises their standards for fashion. Gradually, recommendation systems suggesting products for users have become an indispensable part of e-commerce platforms such as Amazon, eBay, and ShopStyle. These recommendation systems propose products to users based on their preferences and past transaction history [1,2]. Traditional fashion recommendations are carried out by utilizing detailed information about fashion products. Image feature extraction [3,4], hidden information in clothing images, thereby supplementing information for clothing description texts. Subsequently, recognizing the need for clothing product retrieval and proposing combined clothing products tailored to user preferences [5,6].

In the former times, fashion ensembles were often characterized by simplicity, typically comprising only a pair of basic fashion articles such as dresses, skirt-and-top combinations, or tops paired with bottoms, including shorts. However, in contemporary times, fashion ensembles have evolved into significantly more intricate compositions, with a single ensemble potentially incorporating a multitude of fashion items concurrently, such as tops, shorts coupled with jackets, dresses accompanied by jackets and trousers, or tops complemented by skirts. The prevailing sentiment among users is that amalgamating multiple items simultaneously enhances their sartorial appeal. The foremost challenge encountered in such amalgamations lies in ensuring a harmonious convergence of color schemes, patterns, or

* Corresponding author: Bao The Pham

silhouettes among the constituent items, all while preserving a contemporary fashion trends. With such importance, fashion trends have become a topic of interest for many researchers. Zhao et al [7] introduced the Neo-Fashion system, which forecasts fashion trend by data analysis from catwalk images. The NeoFashion design incorporates three modules: data collection and labeling, image segmentation, and trend prediction. The studies have yielded valuable results in trend prediction; however, they have yet to leverage user data. An et al [8] forecast coat trends by leveraging text mining and semantic analysis of fashion blog posts from users.

In this proposal, we construct a fashion recommendation system ensuring the coherence between fashion items within the same outfit suggestion and fashion trends. Our system comprises two stages: fashion apparel detection and similar fashion image retrieval. By utilizing photos and videos from prominent Key Opinion Leaders (KOLs) /influencers on social media platforms as fashion objects, fashion trends can be updated as swiftly as possible. This approach, though straightforward, yields high efficiency as these individuals serve as trendsetters, easily garnering fashion acclaim from users. Our model focuses on 10 item fashion objects: tops, jackets, shorts, pants, dresses, skirts, sunglass, bags, hats, and shoes. In addition to accuracy, we also pay special attention to speed factors to enhance user experience The contributions of the proposal can be summarized as follows:

- Dataset: We collected image data of fashion objects from various e-commerce websites. Additionally, images/videos featuring KOLs/influencers were gathered from popular social media platforms. All experiments were conducted using these self-collected datasets.
- Fashion Apparel Detection: We evaluated the effectiveness of different versions of the YOLOv8 model for the task of detecting fashion objects. The selection of the suitable model was based on accuracy and speed.
- Fashion Image Retrieval: The proposed fashion image retrieval model, our design consists of two CNN branches tasked with extracting image features and producing corresponding results. Subsequently, a voting and ranking mechanism was employed to determine the top K images.

2. Fashion apparel detection

Fashion is one of the fields rich in image data. In there, the primary source of fashion image data is e-commerce websites. To leverage this strength, most fashion-related features are developed by researchers based on images. "Fashion Apparel Detection" is a fundamental problem used to build several common features such as search, classification, and fashion recommendations.

In recent years, with the explosion of deep learning applications, this problem has also been approached by researchers in that direction. Yang, Ming, and Kai Yu [9] proposed a system for detecting real fashion objects using a traditional machine learning model: Linear SVM, combined with the Histogram of Oriented Gradients (HOG) algorithm. Their study focused on 8 categories: suit (top), suit (bottom), shirt, T-shirt, jeans, short pants, short skirt, and long skirt. The system achieved a precision ranging from 45% to 90.3% with a detection speed of up to 20 FPS. The results indicate that the accuracy of traditional machine learning techniques is still limited. After the introduction of the HOG algorithm and its application in various fields, researchers continued to develop the RCNN family of models to address the limitations of the HOG algorithm. Upon its introduction, R-CNN quickly demonstrated its power by achieving high accuracy in object detection tasks. Lao, Brian, and Jagadeesh [10] utilized the R-CNN model for fashion detection. The model achieved 91.25% accuracy during phase-one training and 93.4% during phase-two training. However, for small objects such as shoes and belts, the model required significantly more time for detection. Kucer, Michal, and Naila Murray [11] employed the Mask R-CNN architecture with a Feature Pyramid Network (FPN) backbone for the item detection step in a fashion image query system. The results of the test set showed that the model achieved an overall average precision (AP) of 0.893. Two-stage object detection algorithm, can achieve high accuracy but at the cost of increased computational time. This is a notable limitation of the R-CNN models.

In 2015, the YOLO model [12] was introduced by Joseph Redmon. YOLO is a one-stage object detection algorithm that operates on the principle of "You Only Look Once." The YOLO model addresses the time inefficiencies of R-CNN, reducing computation time while maintaining object detection accuracy. Zheng, Zhihua et al. [13] proposed the YOLOv2-opt model, which is an optimized version of YOLOv2, tailored for the task of fashion object detection. The 5 categories targeted by the authors are: trousers, skirts, jackets, T-shirts, and handbags. Their research demonstrated that the YOLOv2-opt model achieved a mean Average Precision (mAP) of 0.839 and a detection speed of 56ms, outperforming the results of YOLOv2 [14]. Lee, Chu-Hui, and Chen-Wei Lin [15] proposed YOLOv4-TPD, which based on YOLOv4 architecture and the characteristics of transfer learning for the task of fashion object detection, focusing on 5 categories: jackets, T-shirts, pants, skirts, and bags. Their results showed that the proposed model achieved better accuracy for

fashion images with complex backgrounds, which had been a limitation for previous models. Experiments indicated that YOLOv4-TPD outperformed previous: Mean Average Precision (mAP) of 96.01% and a detection speed of 15.633ms.

3. Fashion image retrieval

Fashion image retrieval is also a prevalent feature in e-commerce websites. In the past, product search functionality was often constructed through text comparison. Users searched for desired products by describing the product name or details. However, this method encountered many limitations due to the ambiguity of sentence meanings. Therefore, as the field of deep learning has advanced, researchers have approached this problem using deep learning models. Retrieving desired fashion images from a large dataset is a challenging task due to the measures of similarity and retrieval speed.

Wang, Zhonghao, et al. [16] developed the Visual Attention Model (VAM) architecture to extract attention feature maps from images. The proposed approach improved the accuracy of the top 20 results by 15.9%. Bojana Gajic and Ramon Baldrich [17] propose a fashion image retrieval model based on user's images. The authors propose a model with a structure consisting of 3 Siamese network streams and use Triplet loss for training task. Experiments show that the proposed model achieves a higher effectiveness of 25% compared to the DeepFashion and 41% higher on the DARN dataset. Kinli, Furkan, Baris Ozcan, and Furkan Kirac [18] improve the Capsule Network architecture based on Triplets to learn similar features from three images. The proposed model pays special attention to extracting image features by using stacked convolutional layers or residual connected convolutional layers. The proposed architecture reduces the number of parameters by half while maintaining similar accuracy.

The proposals uniformly adopt the paradigm of tackling the challenge of image retrieval through the utilization of convolutional neural network (CNN) architectures for feature extraction, followed by similarity measures. Notably, feature maps stand out as a pivotal stage, exerting considerable influence on the ultimate query outcomes. A precise characterization of image features enhances the efficacy of the retrieval process, facilitating more accurate and meaningful query results.

4. Proposed method

4.1. Overall Proposed

In this study, we construct an end-to-end system comprising two stages to provide users with fashion recommendations from KOLs/influencers on social media platforms. These individuals consistently exhibit diverse styles, staying abreast of the latest fashion trends and catering to user preferences. In general terms, Fig 1 presents a recommendation system consisting of two stages.

- Fashion Apparel Detection: Detection of fashion objects from input images, users could select their desired fashion items for recommendation.
- Fashion Image Retrieval: Providing top K image suggestions from KOLs/influencers containing similar fashion products to the selected input product. Additionally, we offer article URL information for users interested in exploring additional styles.

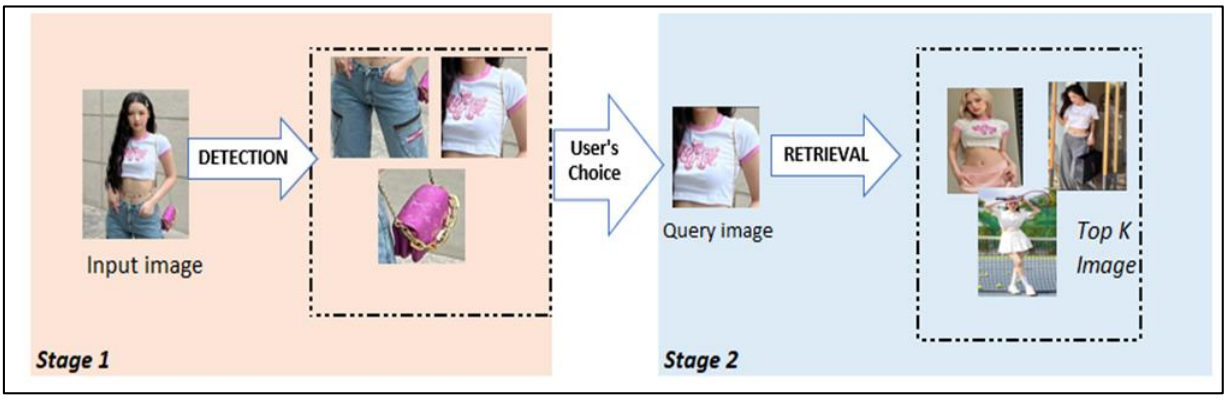

Figure 1 The proposed trend-oriented fashion recommendation system comprises two stages

With the proposing approach, we proceed to collect separate datasets for these two tasks. The available fashion datasets provided free of charge largely consist of simple images that do not align with the images used by users. This discrepancy may result in poor performance of the fashion object detection model. Another reason is that this study focuses only on 10 categories: shirts, jackets, pants, shorts, shoes, glasses, bags, hats, dresses and skirts. The available datasets contain a large number of classifications, which may require preprocessing. Therefore, collecting data from several e-commerce websites for training the fashion object detection model is necessary. Similarly, the image dataset of KOLs/influencers for query purposes is not available, so the data will be collected from popular social media platforms.

4.2. Stage 1: Fashion Apparel Detection

We provide apparel detection fashion method evaluated based on the criteria of accuracy and speed of the model. This is the initial stage in the proposal, it assists users in easily and accurately selecting the desired fashion products as the basis for recommendations in the subsequent stage. Additionally, when considered individually, apparel detection fashion is also a fundamental and essential task for any e-commerce platform.

4.2.1. Data Preparation

In this stage, data preparation is divided into two smaller stages. First, the training process of the object detection model by CNN requires information about ground truth boxes and corresponding object class labels. This information is stored in text files, with the first component being the class name of the object (annotation by numbers from 0-9), followed by four components [x, y, w, h] representing the coordinates of the ground truth boxes. Throughout the training process, the model utilizes information from these stored files to optimize its accuracy. In stage two, we preprocess the images by normalizing them and dividing them by the mean of the entire training dataset to ensure that the input data falls within a defined range. Finally, we obtain complete data for training the fashion object detection model. The dataset is organized and stored in COCO dataset format, consisting of three parts: training set, validation set, and test set. Each dataset includes images and text format label files. The training set and validation set are selected from the collected dataset. To evaluate the model's performance, we use the "Colorful Fashion Dataset" [19] provided freely.

4.2.2. YOLOv8 Architect

According to our understanding, although two-stage object detection models like R-CNN exhibit high accuracy, the large number of parameters significantly reduces detection speed. Meanwhile, YOLO can overcome this limitation while still ensuring accuracy. YOLOv8 [20] is the latest version of the YOLO model family and operates on a single stage object detection, using only a single pass of the input image to predict the presence and location of objects. To meet the requirements of both accuracy and detection speed, we propose using the YOLOv8 architecture for the task of fashion object detection. The architecture is shown in Fig 2.

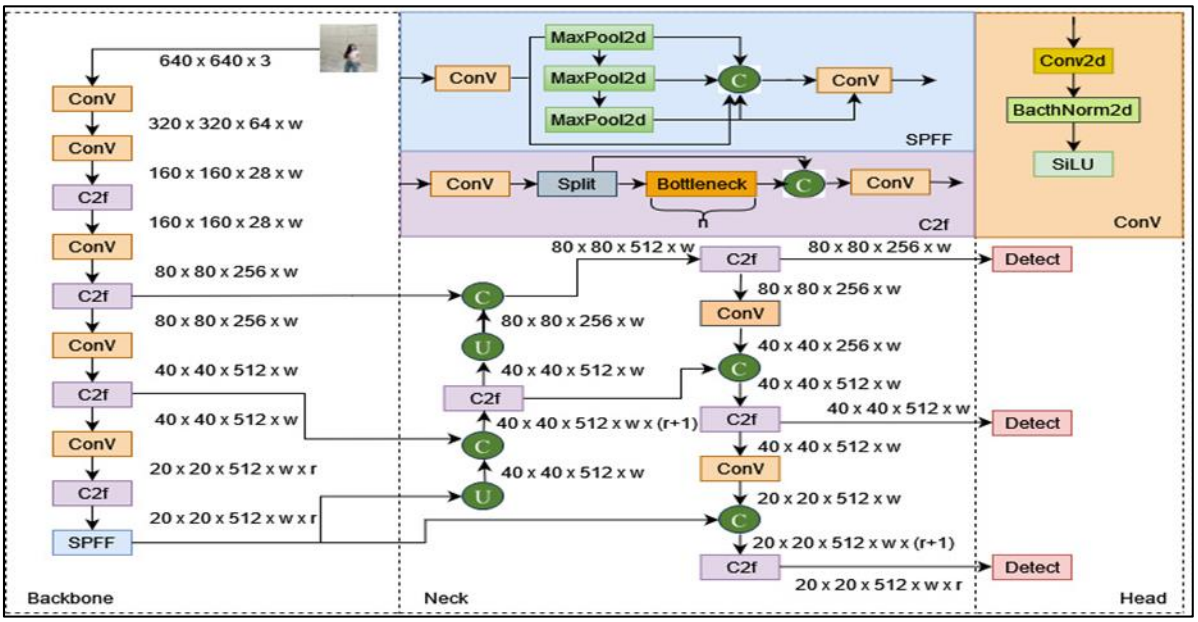

Figure 2 YOLOv8 Architect

Backbone

The task involves extracting image features through convolutional layers. YOLOv8 supports various backbones, including EfficiencyNet, ResNet, and CSPDarkNet. However, as indicated by the results presented in YOLOv5 [21], CSPDarkNet demonstrated the highest efficiency for object detection tasks. Therefore, we propose using CSPDarkNet as the backbone for the architecture of YOLOv8

Neck

This component serves as the connection between the backbone and the head, enhancing the representation of image features extracted by the backbone. To achieve this, YOLOv8 concurrently employs both the Path Aggregation Network (PAN) [22] and Feature Pyramid Network (FPN) [23] architectures. This dual approach allows the model to access information bidirectionally, both top-down and bottom-up, thereby preserving information from both shallow and deep layers.

Loss function

Binary Cross Entropy (BCE) [24]: For the difference between the predicted probabilities and the ground truth. The formula is shown in Equation (1).

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \ldots\ldots\ldots\ldots\ldots(1)$$

Complete Intersection over Union (CioU): Introduced in 2020 [25], plays a crucial role in the regression aspect of object detection. By considering not only the spatial overlap but also the shape discrepancy between predicted and ground truth bounding boxes, CIoU enables more precise localization of objects, leading to improved detection performance, especially in scenarios with tightly packed or overlapping objects. The CIoU loss function is shown in Equation (2).

$$L_{CIoU} = [1 - IoU + \frac{p^2(b,b^{gt})}{c^2} + av] \ldots\ldots\ldots\ldots\ldots(2)$$

Distribution Focal Loss (DFL): Introduced in recent literature [26, 27], addresses the significance of minority class distributions and emphasizes the model's predictive adequacy regarding the real-world class distributions. DFL incorporates a focal mechanism that dynamically adjusts the loss contribution from each instance based on its classification difficulty, thereby enhancing the model's ability to effectively learn from imbalanced datasets and improve performance in minority classes. The formula used to work as shown in Equation (3).

$$DFL_{(S_i,S_{i+1})} = [-((y_{i+1} - y)\log(S_i) + (y - y_u)\log(S_{i+1}))] \quad \ldots\ldots\ldots\ldots\ldots[3]$$

4.2.3. Training and Evaluation

First, the model will be trained using data collected from various e-commerce platforms, and hyperparameters will be optimized through a validation set. Subsequently, the model's effectiveness will be evaluated using the "Colorful-Fashion" [19]. The work is shown in Fig 3.

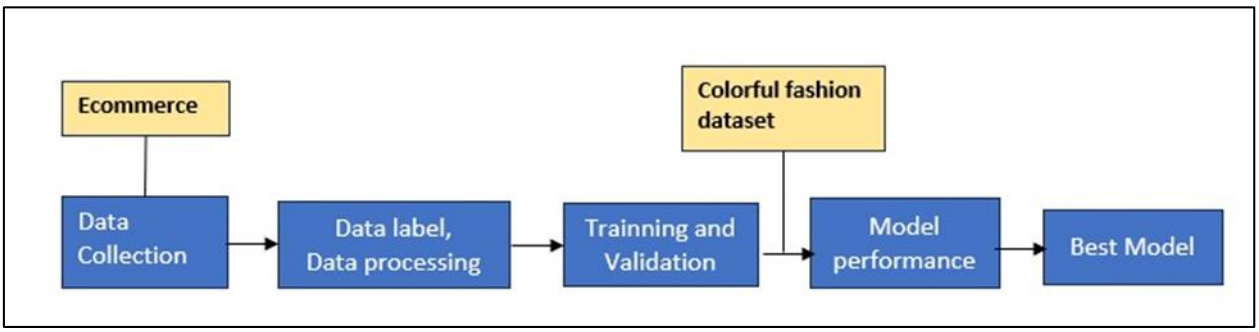

Figure 3 Step of proposed method for Fashion Apparel Detection

The training dataset in the preparation phase is utilized for learning, while the validation dataset is employed for parameter tuning. During model training, the thesis sets the input image size to 640x640, as recommended by YOLOv8. The batch size is set to 64 to ensure the model can process a sufficient number of images in each training iteration.

Hyperparameters such as the optimizer, initial learning rate (lr0), final learning rate (lrf), momentum, and decay are initialized with the default values provided by YOLOv8. Additionally, the mosaic augmentation technique is applied during the last 10 epochs. It is worth noting that excessive use of mosaic augmentation, as indicated by original studies, can lead to slow model convergence and reduced accuracy.

The model is evaluated based on the accuracy of detecting and classifying fashion objects. The metrics utilized to evaluate the model include Average Precision (AP), Mean Average Precision (mAP) and F1 Score.

Average Precision (AP): AP measures the average accuracy of the model in detecting objects within each class. The mathematical formula for the AP is shown in Equation (4).

$$AP = \int_0^1 Precison(Recall)d(Recall) \ldots\ldots\ldots\ldots (4)$$

Mean Average Precision (mAP): mAP calculates the average value of AP across all classes. If there are k categories, mAP is computed as the average AP over all categories. The mathematical formula is denoted in Equation (5)

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \ldots\ldots\ldots\ldots\ldots (5)$$

F1 Score: The F1 score is a metric that harmonizes Precision and Recall. It provides a balanced measure of the model's performance in terms of both precision and recall. Equation (6) denotes the F1 score.

$$F1 = \frac{2.Precision.Recall}{Precison+Recall} \ldots\ldots\ldots\ldots\ldots (6)$$

4.3. Stage 2: Recommend the top K outfit for users

4.3.1. Data Preparation

The image data is collected from two popular social media platforms: Instagram and TikTok. Images from Instagram are saved and videos gathered from TikTok are processed by extracting frames every 3 seconds. After saving, they are passed through the fashion object detection model to detect fashion products. Subsequently, feature extraction is performed, and the resulting vectors are stored as indices. Additionally, to enhance user experience, we store some post information for accessing KOLs/Influencer post URLs, thus expanding the proposed style recommendations for users.

4.3.2. Proposed method

Our proposal consists of two parallel branches of CNN architectures, each producing different streams of results. Subsequently, we construct a voting and ranking algorithm to determine the top K output images for the model. With this approach, we aim to achieve highly reliable results for input image queries. The proposed method is shown in Fig 4.

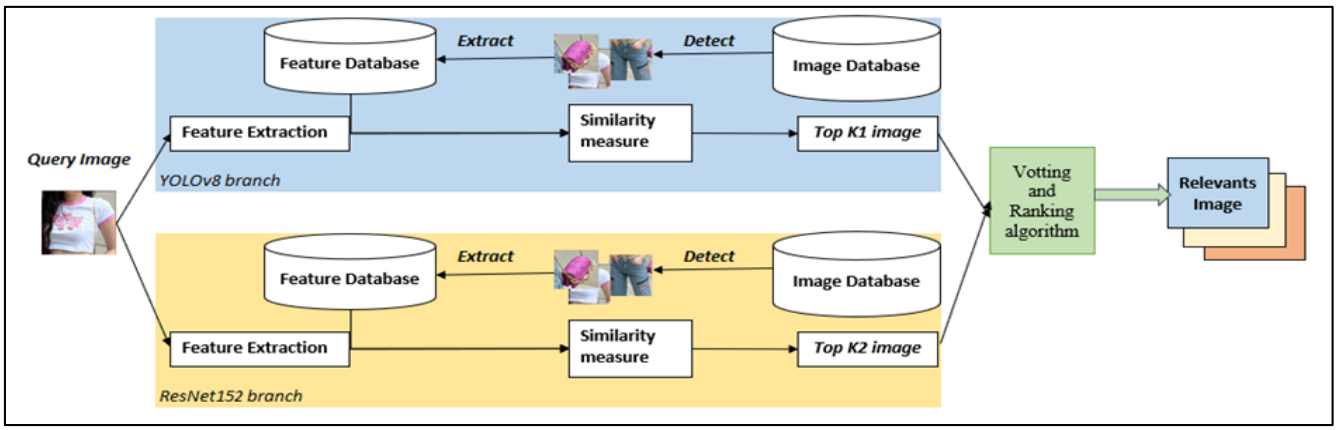

Figure 4 Proposed Method For Fashion Image Retrieval

ResNet152 is a deep neural network introduced by Microsoft Research in the paper "Deep Residual Learning for Image Recognition" (He et al., 2016) [28]. It consists of 152 convolutional layers with residual blocks and has been pre-trained

on the large-scaled ImageNet dataset. Furthermore, ResNet models have been widely used and proven successful in various image retrieval tasks [17, 29, 30]. Therefore, in the initial phase, we employ the ResNet152 architecture to extract image features.

In the other branch, we utilize the CNN architecture of the pre-trained YOLOv8 model as the foundation. This architecture has been successfully trained and incorporates sophisticated features learned from "Fashion Apparel Detection" in stage one. They combining ResNet152 and YOLOv8 architectures, our proposed model aims to capture both high-level semantic features and fine-grained details of fashion images, thereby enhancing the accuracy and effectiveness of fashion image retrieval.

In YOLOv8n branch, the feature map has dimensions of 40 x 40 x 74. It converts a vector into one-dimensional array (74,). This refined vector is then stored in the index structures managed by the Facebook AI Similarity Search (FAISS) library, facilitating efficient retrieval and comparison. In ResNet152 branch, the feature map with significantly larger dimensions, specifically (2048). These feature vectors are also indexed using FAISS. This dual-indexing approach leverages the strengths of both YOLOv8n and ResNet152, enhancing the model's overall capability for precise and efficient image retrieval in fashion related applications.

Feature vectors are searched based on the L2 distance. This distance metric indicates similarity by measuring how close the feature vectors in the database are to the feature vector of the input image. The smaller the L2 distance, the more similar the vectors are, and the larger the distance, the less similar they are. The L2 distance is denoted in Equation (7).

$$d_2(Q,T) = \sqrt{\sum_{i=0}^{N=1}(Q_i - T_i)^2} \text{.....(7)}$$

Speed is one of the challenges when accessing data, especially as the size of the vector database increases, leading to slower query results. To address this limitation, we propose the Facebook AI Similarity Search (FAISS) library. FAISS operates on the basis of an indexing mechanism, where it stores the feature vectors under indices and then performs computational tasks and searches.

By employing this dual approach of voting and ranking, our algorithm effectively synthesizes the results obtained from the feature vectors for retrieved image. Feature vectors that receive higher vote are prioritized and ranked more favorably, indicating a stronger among the retrieved vectors regarding their relevance to the query.

Algorithm: Generate Output
Input: Top K_1 neigh image in YOLOv8 branch Top K_2 neigh image in ResNet 152 branch
Output: Top K_{neigh} image
Function Voting and Ranking (K1, K2)
Candidates <- K_1 U K_2
Vote <- $K_1 \cap K_2$
Subset <- DESC(Candidates\Vote)
Queue <-Vote U Subset
Get K_{neigh} top K images from Queue

4.3.3. Evaluation

To evaluate the effectiveness of our model, we utilize the top-K measure defined as Equation (8). The top-K measure assesses the system's retrieval capability by considering the proportion of similar images within the top-K recommended items, where K is a predefined threshold. This threshold is determined based on practical considerations or specific requirements.

$$P@K = \frac{\sum hit(q,K)}{K} \text{.....(8)}$$

Where *hit(q, K) = 1* if the returned image is similar to the query image, where K denotes the number of initial image results that the system returns.

In our proposal, we evaluate the system's performance using top K with K=5 and K=10. These values provide insights into how well the system performs in recommending relevant images within the top 5 and top 10 ranked items.

5. Experimental result

We conducted experiments with five different versions of the YOLOv8 model to identify the most suitable model for the fashion object detection stage of our recommendation system.

To demonstrate the authenticity of the experiment, we used mean Average Precision (mAP) at thresholds ranging from 50 to 95, Speed, number of Parameters (Param), and Floating Point Operations per Second (FLOPs) as evaluation metrics. The test results are summarized in Table 1.

Table 1 Experiment result of versions of the YOLOv8 model for the problem Fashion Apparel Detection

Model	Size	mAP (val 50-95)	Speed (CPU) ms	Speed (Tesla V100-SXM2-16GB) ms	Param (M)	Flops (B)
YOLOV8n	640	0.795	69	1.7	3.007	8.1
YOLOv8s	640	0.804	152	2.9	11.129	28.5
YOLOv8m	640	0.806	374.6	5	25.846	78.7
YOLOv8l	640	0.809	732.9	8.3	43.614	164.9
YOLOv8x	640	0.807	1062.6	12.1	68.133	257.4

The results indicate that the YOLOv8n (nano) model achieves high efficiency with a mAP of 0.796 and a processing speed of only 69ms on a CPU. Compared to YOLOv8s, the model sacrifices detection speed for a marginal mAP improvement of 0.9%-1.2%. This trade-off can potentially affect user experience in Fig 5.

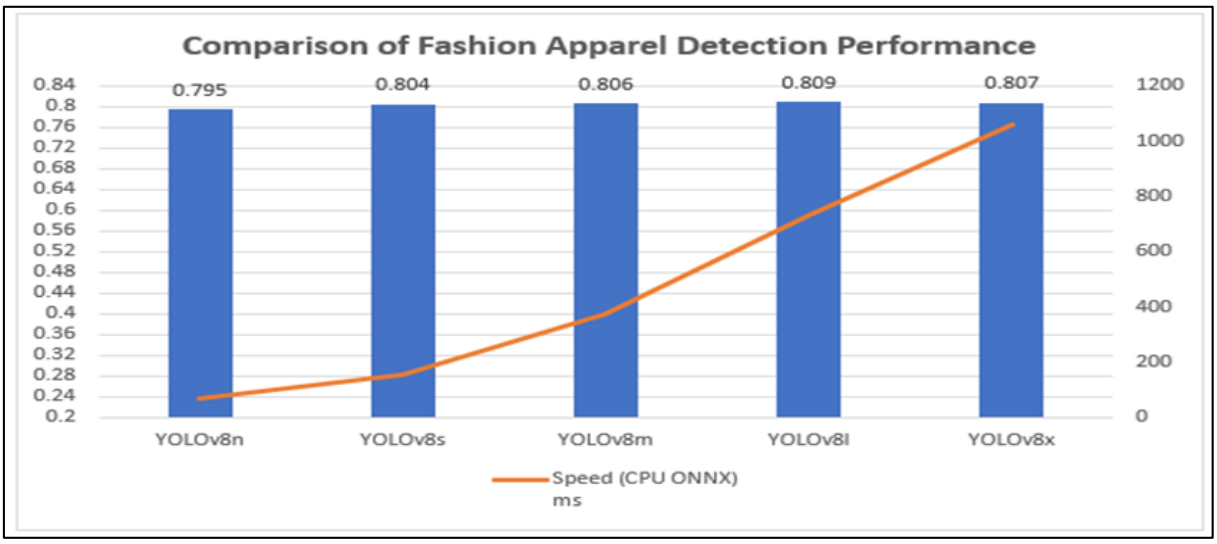

Figure 5 Comparison of Fashion Apparel Detection Model Performance

On the test dataset, the YOLOv8n model achieved a mAP@50 of 0.868, a speed of 114 FPS, and an F1 score of 0.81. Therefore, we propose using the YOLOv8n model for the task of fashion object detection. The high mAP and F1 score, combined with the efficient processing speed, make YOLOv8n an optimal choice for integrating into our fashion recommendation system.

To comprehensively assess the efficacy of our fashion image retrieval model, we curated a test dataset consisting of 100 images, each featuring fashion products that exhibit similarity in classification, color scheme, and stylistic attributes. Furthermore, to test the robustness of our model, we used some noise images. These noise images were deliberately

chosen to deviate significantly from the defined similarity criteria, varying in classification, color palette, and stylistic elements. This deliberate inclusion of noise images serves to challenge the model's validate its capacity to filter out irrelevant or dissimilar fashion items.

For each category: shirt, dress, pants, and hat, we get five samples. Subsequently, we employed evaluation metrics to quantify the performance of our model. Specifically, we assessed the retrieval accuracy of the top 5 and top 10 returned results, measured against the initial similarity criteria established during dataset creation. By scrutinizing the precision of the model's predictions against these predefined criteria, we gain valuable insights into its ability to retrieve similar fashion items. The results are shown in Table 2 and Fig 6.

Table 2 Experiment result of Fashion Image Retrieval

Sample	Attribute	P@5	P@10
Query 1	Category: Shirt Color: white Style: T-shirt	90%	80%
Query 2	Category Skirt Color: Black Style: Short	90%	70%
Query 3	Category: Dress Color: Black Style: Long	100%	80%
Query 4	Categoryl Hat Color: Black Style: Beret	90%	90%
Query 5	Category: Pants Color: Blue Style: Long	90%	90%

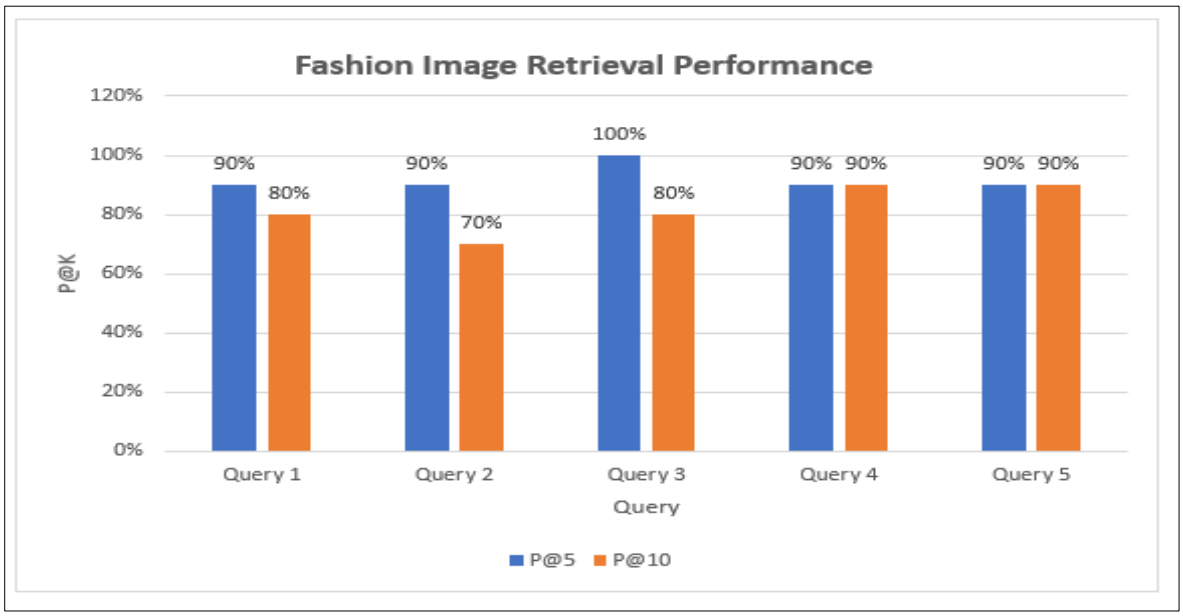

Figure 6 Fashion Image Retrieval Performance

During our experimentation with various samples encompassing both stages of the recommendation system, we noted the system's capability to adeptly fulfill the criteria of accurate classification, style identification, color discrimination,

and trend prediction, leveraging image data sourced from influential KOLs/influencers. The integration of URL post from KOLs/influencers enriches the recommendation system's understanding for users. Leveraging this data, users can capture nuanced style variations and emergent trends, enhancing its capacity to provide personalized and trend-conscious fashion recommendations to users.

Through these experiments, we affirm the viability and effectiveness of our recommendation system in fashion trends. Fig 7 provides sample result of the outcomes achieved by our comprehensive recommendation system.

Figure 7 Experiment result of the proposed system

6. Conclusion

In our research, we propose a novel approach to the recommendation problem within ten fundamental fashion categories: sunglasses, hats, jackets, shirts, pants, shorts, dresses, skirts, bags, and shoes. The proposed consists of two integral sub-models: a fashion object detection model utilizing YOLOv8n and a fashion image retrieval model supported by the FAISS library. Fashion Apparel Detection leverages the YOLOv8n was selected due to its balance between accuracy and computational efficiency, it is suitable for practical applications where both speed and precision are critical. Fashion Image Retrieval sub-model employs a dual-branch CNN architecture for fashion image similarity search. This architecture incorporates two parallel CNN branches: one reuses the CNN network from the fashion object detection model, while the other utilizes a pretrained ResNet152 architecture. The results from these two branches are combined using a voting and ranking function to increase the reliability of the output, ensuring that the most relevant results are prioritized. End-to-end the system, our results demonstrate that the proposed model achieves a high level

of accuracy and processing speed, confirming its viability for real-world applications. The system's performance indicates its potential utility for end-users, providing recommendations with trending fashion.

However, despite these promising results, we identify several limitations and areas for improvement:

- The data from KOLs/ Influencers is limited in quantity, which results in recommendations that lack richness and diversity. Expanding this dataset could enhance the variety and quality of user recommendations.
- The model currently operates based on ten basic fashion items. Extending the dataset to include a wider array of fashion items would provide users with more choices.
- Classification of sunglasses and shoes can be easily confused with the background in images. It is necessary to guide users to provide images that focus on the fashion items they want recommendations for, thereby improving the model's accuracy.

Future development directions for this research include expanding the dataset to encompass additional fashion items and increasing the data volume from KOLs and influencers. To ensure the model stays current with fashion trends, an automated pipeline will be established to remove outdated data and periodically update the dataset with new information. Furthermore, the model will be enhanced by incorporating additional user personal information fields, such as gender, height, weight, and clothing size, to improve the relevance and personalization of the recommendations. Through these efforts, we aim to refine the proposed model, thereby improving its practical applicability and performance in diverse real-world scenarios.

Compliance with ethical standards

Acknowledgments

We would like to extend our heartfelt gratitude to the editor and the reviewers for their invaluable comments and suggestions.

We deeply appreciate their time, effort, and expertise in reviewing our manuscript. Their constructive criticism and thoughtful.

Disclosure of conflict of interest

The authors declare no potential conflict of interest.

References

[1] Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In: , , Springer, 2010. p.1–35

[2] Zhang S, Yao L, Sun A, Tay Y. Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR). 2019;52(1):1–38.

[3] Wang XY, Zhang BB, Yang HY. Content-based image retrieval by integrating color and texture features. Multimedia tools and applications. 2014;68(3):545–569.

[4] Rahul K, Agrawal R, Pal AK. Color image quantization scheme using DBSCAN with K-means algorithm. In: Springer. 2014. p.1037–1045.

[5] Ping Tian D, others. A review on image feature extraction and representation techniques. International Journal of Multimedia and Ubiquitous Engineering. 2013;8(4):385–396.

[6] Navneet D. Histograms of oriented gradients for human detection. International Conference on Computer Vision & Pattern Recognition. 2015;2(0):886-893

[7] Zhao L, Li M, Sun P. Neo-fashion: A data-driven fashion trend forecasting system using catwalk analysis. Clothing and Textiles Research Journal. 2024;42(1):19–34.

[8] An H, Park M. Approaching fashion design trend applications using text mining and semantic network analysis. Fashion and Textiles. 2020;7(1):34.

[9] Yang M, Yu K. Real-time clothing recognition in surveillance videos In: , ,IEEE. 2011. p.2937–2940.

[10] Lao B, Jagadeesh K. Convolutional neural networks for fashion classification and object detection. CCCV 2015 Comput. Vis. 2015;546(0):120–129.

[11] Kucer M, Murray N. A detect-then-retrieve model for multi-domain fashion item retrieval. In: , ,Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019. p.0–0.

[12] Han X, Chang J, Wang K. You only look once: unified, real-time object detection. Procedia Computer Science. 2021;183(1):61–72

[13] Feng Z, Luo X, Yang T, Kita K. An object detection system based on YOLOv2 in fashion apparel. In: , , IEEE. 2018. p.1532–1536.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: , ,Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p.7263–7271.

[15] Lee CH, Lin CW. A two-phase fashion apparel detection method based on YOLOv4. Applied Sciences. 2021;11(9):3782.

[16] Wang Z, Gu Y, Zhang Y, Zhou J, Gu X. Clothing retrieval with visual attention model. In: , ,IEEE. 2017. p.1–4.

[17] Gajic B, Baldrich R. Cross-domain fashion image retrieval. In: , , Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018. p.1869–1871.

[18] Kinli F, Ozcan B, Kirac F. Fashion image retrieval with capsule networks. In: , , Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019. p.0–0.

[19] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, & S. Yan Fashion Parsing with Weak Color-Category Labels [Internet]. Baltimore:IEEE; © 2013 [cited 2013 Sep 01]. Available from https://sites.google.com/site/fashionparsing/home.

[20] Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. 2023.

[21] Jocher, Glenn, et al. Ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo. 2022.

[22] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: , , Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p.8759–8768.

[23] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: , , Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p.2117–2125.

[24] Goodfellow I, Bengio Y, Courville A. In: , , Deep learning. MIT press, 2016.

[25] Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the AAAI conference on artificial intelligence. 2020;34(07):12993–13000

[26] Li X, Wang W, Hu X, Li J, Tang J, Yang J. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In: , , Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021. p.11632–11641.

[27] Li X, Lv C, Wang W, Li G, Yang L, Yang J. Generalized focal loss: Towards efficient representation learning for dense object detection. IEEE transactions on pattern analysis and machine intelligence. 2022;45(3):3139–3153.

[28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: , , Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p.770–778

[29] Ibrahimi S, Noord vN, Geradts Z, Worring M. Deep metric learning for cross-domain fashion instance retrieval. In: , , Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019. p.0–0.

[30] Morelli D, Cornia M, Cucchiara R, others . FashionSearch++: Improving consumer-to-shop clothes retrieval with hard negatives. CEUR Workshop Proceedings.2021;2947(0):0-0