Multimedia Tools and Applications (2025) 84:40195–40214
https://doi.org/10.1007/s11042-025-20753-x

OutfitAI: shop the outfit with a deep learning-based intelligent expert system

Emanuele Balloni[1] · Rocco Pietrini[1] · Emanuele Frontoni[2] · Adriano Mancini[1] · Marina Paolanti[2]

Received: 11 April 2024 / Revised: 12 September 2024 / Accepted: 5 March 2025 /
Published online: 20 March 2025

Abstract
In an age where consumer preferences are as diverse as they are dynamic, the ability to offer personalized fashion recommendations at scale remains a significant challenge for retailers. Consumers seek a shopping experience that not only understands their unique style preferences but also dynamically adapts to their evolving tastes. The fashion industry is at a crossroads, facing increasing consumer demand for personalization, sustainability and transparency in a rapidly evolving digital marketplace. Traditional retail practices, while rich in tradition and artistry, often struggle to up-to-date with the rapidly, ethically-conscious and technology-driven expectations of today's consumers. "OutfitAI" is designed to address these challenges by leveraging the power of deep learning to revolutionize the fashion retail experience. By automating the process of background removal in fashion images, using advanced algorithms for personalized product matching, and integrating sustainability filters into the product discovery process, OutfitAI aims to deliver a shopping experience that is not only personalized and engaging, but also aligned with the ethical and environmental values of the contemporary consumer. Unlike existing solutions, OutfitAI uses state-of-the-art semantic segmentation for precise background removal, enabling detailed feature extraction from fashion images. This process enables accurate matching of user-uploaded images with similar fashion items from an extensive database of eco-friendly and ethically produced products sourced from leading e-tailers. Setting itself apart from the current state of the art, OutfitAI places a strong emphasis on ethical data use and privacy, implementing robust measures to ensure user privacy and transparency. It also pioneers the integration of sustainability into the digital fashion discovery process, promoting responsible consumption patterns among users. Through a comprehensive system architecture that combines technical innovation with a commitment to ethics and sustainability, OutfitAI not only addresses the technological needs of the fashion retail industry, but also responds to the growing demand for more responsible and transparent consumer technologies.

Keywords Background removal · FashionAI · Transformers

✉ Emanuele Balloni
e.balloni@pm.univpm.it

[1] Department of Information Engineering (DII), Università Politecnica delle Marche, 60131 Ancona, Italy

[2] Department of Political Sciences Communication and International Relations, University of Macerata, 62100 Macerata, Italy

1 Introduction

In recent years, it has become common practice for shoppers to familiarize themselves with other people's fashion outfits through online retail platforms and e-tailers, leading to their own efforts to acquire fashion items. In addition to general-purpose online marketplaces and e-tailers are widely used. The tracking of full-body outfit visuals on these platforms' applications or e-commerce sites plays a pivotal role in enhancing shoppers' fashion-oriented shopping experiences, influencing their fashion choices and decision-making processes [30]. Today's consumers not only want personalized shopping experiences that match their unique style preferences, they're also demanding that these experiences align with their growing concerns about sustainability and transparency. This shift underscores the urgent need for innovation that not only understands and adapts to individual tastes, but also promotes ethical practices and environmental sustainability within the fashion industry. Against this scenario, the critical role of advanced technology solutions, particularly in the area of image processing and background removal, becomes evident [3].

The importance of sophisticated background removal techniques in meeting these consumer demands cannot be overstated. Accurate and efficient background removal is fundamental to the creation of high-quality, visually appealing product images, which are essential for personalisation algorithms to work effectively. Background removal is a crucial image processing technique used extensively in the fashion industry, especially in the e-commerce sector. In online shopping, customers rely heavily on product images to make informed purchasing decisions. The quality and accuracy of the visual experience can significantly impact their buying decisions. Removing the background from product images not only eliminates distractions but also helps to highlight the product's key features, thus making it easier for customers to evaluate the product and make a purchase [21]. In the fashion domain, background removal can significantly improve the visual appeal of product images [36]. When consumers browse e-commerce websites, they are looking for high-quality images that accurately represent the products they are interested in. By removing the background and isolating the product, the image becomes more visually appealing and professional-looking. This is especially important in the fashion industry, where the appearance of products is a key factor in purchasing decisions [2]. Background removal can also help to standardize product images, easing for consumers to compare different products and make informed decisions. Overall, background removal is an essential technique for creating high-quality, professional-looking product images in the fashion domain. Although background removal can be done manually, it is a time-consuming and tedious task, especially when dealing with large volumes of images [1, 19]. Therefore, automating this process through deep learning has become a popular approach in recent years. Artificial Intelligence (AI) and deep learning techniques have been increasingly used in various image processing tasks, and semantic segmentation is the most promising method for background removal [9, 37]. Semantic segmentation can be applied to background removal by training a neural network on a large dataset of images with annotated foreground and background regions. The resulting model can then be used to remove the background from new images. In the fashion industry, several approaches to background removal have been proposed [23, 26]. However, due to the nature of fashion images, which often contain complex patterns and textures, it remains a challenging task.

To address this issue, in this paper it is proposed OutfitAI an expert system that employs specialised deep learning-based strategies tailored to the fashion context. Our research includes techniques such as unsupervised semantic segmentation by contrasting object mask proposals and cross-image pixel contrast for semantic segmentation. In addition, OutfitAI

incorporates transformer-based neural networks, adapting these powerful models to refine the process of isolating fashion items from their backgrounds. These approaches not only promise to improve the accuracy and efficiency of background removal in fashion images, but also play a critical role in OutfitAI's mission to revolutionise the fashion retail industry through a deep-learning enhanced outfit-based shopping experience. These approaches include Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals [34], Cross-Image Pixel Contrast for Semantic Segmentation [35], and two neural networks based on transformers [32, 38].

In particular, the choice of using transformers arises from they ability to adapt at handling variations in scale, orientation, and appearance within fashion images. They can effectively learn spatial relationships and understand the semantic meaning of different regions in an image, enabling precise and reliable background removal. Fashion datasets, in fact, can be vast and diverse, requiring models with robust generalization capabilities. Transformers, with their self-attention mechanisms and parallel processing capabilities, can effectively leverage the abundant data available for training, leading to improved performance and accuracy. By incorporating transformers into the background removal process, fashion industry professionals can benefit from automated and efficient background removal techniques that yield high-quality results. These techniques enhance the visual experience for customers by ensuring that fashion products are presented in an appealing and contextually appropriate manner. Therefore, the use of transformers in background removal techniques represents a significant advancement in the fashion industry, enabling accurate and efficient processing of fashion images while maintaining the highest standards of visual quality.

At its core, the platform features an easy-to-use interface, accessible through both web and mobile applications, which makes it easy for users to navigate and interact with the platform to discover fashion that matches their personal preferences. This enables precise background removal, setting the stage for the deep learning and recommendation engine to extract and analyse features of fashion items with unparalleled accuracy. OutfitAI matches these features against a comprehensive product database sourced in real time from leading e-tailers to ensure recommendations are both personalised and up-to-date, using transformer-based neural networks. The architecture is supported by a robust cloud infrastructure that provides scalable computing resources and secure data management, while a continuous feedback loop with analytics ensures that the system evolves in line with user preferences and emerging fashion trends. Through this advanced integration of technology and user-centric design, OutfitAI offers a revolutionary approach to outfit-based shopping, making it an indispensable tool for fashion enthusiasts and a paradigm of innovation in retail technology. Beyond technical capabilities, OutfitAI is committed to ethical data use, offering robust privacy protections and user control over personal data, in contrast to less transparent practices in the industry. Sustainability is built into the structure of its recommendation engine, which prioritises eco-friendly and ethically produced fashion items, highlighting OutfitAI's commitment to responsible retail.

The main contributions of this paper can be summarized as follows: i) the development of an innovative semantic segmentation approach for precise background removal in fashion images, ii) the application of transformer-based neural networks for detailed feature extraction, iii) the integration of a scalable, cloud-based architecture that combines real-time e-tailer data with a sophisticated recommendation engine, iv) a highly personalised shopping experience, bridging the gap between fashion enthusiasts and the latest market trends. v) the discussions of ethical considerations of user data and the promotion of sustainable fashion, providing a comprehensive roadmap for future applications at the intersection of AI, deep learning and e-commerce.

The paper is organised as follows. Section 2 provides an overview of state-of-the-art approaches for background removal. Section 3 presents OutfitAI system architecture. Section 4 presents the experimental results and the corresponding analysis. We compare the proposed method with existing techniques and demonstrate its effectiveness in removing backgrounds from fashion images. Finally, in Section 6, we conclude the paper and discuss potential future directions for improving the proposed method.

2 Related works

Background removal is a well-established technique in computer vision and deep learning, with origins in some of the earliest and most recognised applications in intelligent visual surveillance systems [10]. Designed to monitor human activity, these systems have played a key role in areas such as traffic monitoring on roads, and security and operational monitoring in airports and maritime environments [18]. The core functionality of these systems often relies on the ability to distinguish and isolate moving subjects or objects of interest from their static or dynamic backgrounds [4]. This fundamental process enables the effective analysis of behaviours, patterns and incidents within the monitored environments, demonstrating the critical role that background removal has played in the development and operational success of surveillance technology since its inception [5].

In contrast to its established role in literature, the technique of background removal has not been as widely recognised or extensively explored in the fashion domain. Despite its potential to significantly enhance the analysis and presentation of fashion imagery by isolating garments from distracting backgrounds, this application has remained relatively under-exploited. The fashion industry, with its emphasis on aesthetics, styling and visual appeal, could greatly benefit from the integration of background removal technologies. Such techniques could improve the clarity and focus of fashion imagery, facilitating better recognition and analysis of styles, trends and garment features. However, the specific use of background removal in fashion imagery represents an area of interest that has not yet reached the same level of prominence or development as its applications in more traditional computer vision domains. This section provides a comprehensive review of the existing literature in the field of fashion, focusing specifically on studies that employ deep learning techniques for background removal from fashion images. The aim is to assess how these methods improve the analysis and interpretation of fashion data by isolating garments from their various backgrounds. By examining the progress and results of these pioneering studies, we aim to highlight the technological advances and methodological approaches that have significantly contributed to the refinement of automated fashion analysis.

Liang et al. [22] explored the process of extracting background from fashion photographs to improve data integrity and the effectiveness of computational models. Using images that clearly depicted feature people dressed in garments, they applied salient object detection techniques to successfully separate fashion items from their backgrounds in accordance with their objectives. These processed images, known as 'rembg' images, are distinct from their original counterparts in the dataset. Their research involved a thorough series of comparative analyses between the original and the background-removed images across different dimensions of model training. These dimensions include different model architectures, initial setup procedures, integration with different training extensions and data augmentation strategies, and applicability to different types of tasks.

AFRIRAZER emerges as a key deep learning model dedicated to selectively removing background and skin from traditional African fashion images [24]. The unique challenge here lies in the under-representation of traditional African images in standard fashion datasets. This initiative is a critical component of a broader project that aims to create a curated, anonymized and relevant collection of traditional African fashion images for use in artificial intelligence applications. AFRIRAZER used deep learning algorithms to extract fashion elements from their backgrounds, taking into account the specific nuances and textures that characterize African clothing.

Building on the foundation of innovative methods such as AFRIRAZER, in [23], the authors introduced the WEARStyle dataset, a large-scale collection tailored for fashion analysis, and implemented two techniques aimed at refining image quality by eliminating extraneous pixels. These methods include a Single Shot MultiBox Detector (SSD) based approach for human figure detection and a Pyramid Scene Parsing Network (PSPNet) based strategy for precise pixel selection. With this development, they demonstrated the impact of targeted data enhancement and sophisticated pixel filtering on improving the accuracy of fashion image analysis.

In the context of social media, Kinli et al. [20] proposed the Instagram Filter Removal Network (IFRNet), an approach designed to neutralize the effects of filters in social media image analysis. IFRNet was based on the notion that applying a filter to an image essentially adds a layer of stylistic information, positioning the challenge as akin to solving a reverse style transfer problem. By normalizing this added style information at each level of the encoder, IFRNet removed the visual effects of filtering.

Compared to state-of-the-art approaches in the field, OutfitAI introduces several key innovations and improvements, particularly in the area of fashion AI. First, it advances the technique of background removal by using deep learning models that are finely tuned to the complex textures and patterns of fashion images, significantly outperforming standard methods in terms of accuracy and efficiency. This underlying technology enables more accurate feature extraction and product matching. OutfitAI also has a strong commitment to ethics and sustainability. While many platforms collect user data with minimal transparency, OutfitAI prioritizes user privacy by implementing strict privacy measures and ensuring that users have full control over their personal information. This approach addresses growing concerns about data security and privacy in the digital age. When it comes to sustainability, OutfitAI goes beyond the norm by incorporating environmental and ethical considerations into its recommendation engine. It not only suggests fashion items based on style and preference, but also filters these recommendations through a sustainability lens, promoting products from brands that adhere to responsible manufacturing practices. This is a significant departure from most current systems, which typically focus on style and trend matching without considering the environmental and ethical implications of their recommendations.

3 Materials and methods

This section describes the overall system architecture of "OutfitAI", the datasets used to train and evaluate the deep learning models for background removal, and the ethical considerations and focus on sustainability that are integral to the development and deployment of the system.

"OutfitAI" is built on a robust, scalable cloud infrastructure designed to process and analyze large volumes of fashion images with unparalleled efficiency. This architecture is

meticulously divided into three primary components, each designed to fulfill a specific role within the system:

- **Background Removal Module**: The core component of OutfitAI uses advanced transformer-based neural networks to achieve accurate background removal from fashion images. By harnessing the power of transformers, the module excels at isolating fashion items from complex backgrounds, enabling more accurate feature extraction and ensuring subsequent product recommendations are as relevant as possible to the user's uploaded content.
- **Recommendation Module**: this module uses sophisticated similarity algorithms to match the segmented fashion items with an extensive range of products from partner e-tailers. By analysing the visual and stylistic features of the extracted items, the Recommender Module identifies and suggests products that closely match the user's preferences, facilitating a personalised shopping journey that reflects the user's tastes and nuances of style.
- **Ethics and Sustainability Module**: Reflecting OutfitAI's commitment to responsible AI development, this module is based on the Ethical Guidelines for Trustworthy AI. It ensures that all operational aspects of OutfitAI, from data handling and privacy to product promotion, adhere to strict ethical standards. It also incorporates sustainability considerations into the recommendation process by prioritising eco-friendly and ethically produced fashion items, encouraging users to make more conscious and sustainable fashion choices.

Figure 1 illustrates the comprehensive architecture of OutfitAI.

3.1 Background removal module

As stated in the Introduction, for this task, we employed Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals, Cross-Image Pixel Contrast for Semantic Segmentation, and two transformer-based neural networks. Each model offers unique features and capabilities that contribute to the effectiveness of the background removal process. These models were chosen for their ability to consistently perform well across various image processing and related tasks in the literature. They have been shown to outperform traditional

Fig. 1 OutfitAI System Architecture. This diagram illustrates the comprehensive architecture of OutfitAI, showing its three core modules. The Background Removal Module uses transformer-based neural networks to accurately segment fashion items from complex backgrounds, ensuring high quality input for feature extraction. The Recommender Module uses similarity algorithms to match these processed images with a wide range of products from partner e-tailers, providing users with personalised fashion recommendations. Finally, the Ethics and Sustainability Module is guided by the ethical guidelines for trustworthy AI, embedding sustainable practices and ethical considerations into every aspect of the system, from data handling to product promotion

methods in tasks such as image classification, feature extraction, and object detection [32, 34, 35, 38].

3.1.1 Transformers networks

Segmentation semantic task has been implemented with two transformers deep learning model using different architectures. Instead of building these models from scratch to solve a similar problem, pre-trained models on similar tasks have been used as a starting point to fine-tune on the iMaterialist dataset. In particular, the architectures taken were pretrained on ADE20K dataset [39]. The Segmenter [32] model is based on a fully transformer-based encoder-decoder architecture, mapping a sequence of patch embeddings to pixel-level class annotations. The sequence of patches is encoded by a transformer encoder and decoded, by either a point-wise linear mapping or a mask transformer. The encoder was built upon the Vision Transformer model (ViT) used for image classification, that employs a Transformer-like architecture over patches of the image previously obtained; among all the existing variants, the following ViTs were considered: "Tiny" and "Base" models [13]. For the decoder a Mask Transformer was used, where the sequence of patch encodings is decoded to a segmentation map. The decoder learns to map patch-level encodings coming from the encoder to patch-level class scores. Next, these patch-level class scores are upsampled by bilinear interpolation to pixel-level scores using a softmax followed by a norm; these scores form the final segmentation map. Segmenter model is trained end-to-end with a per-pixel cross-entropy loss. At inference time, argmax is applied after upsampling to obtain a single class per pixel. SETR is based on a pure self-attention encoder combined with a simple decoder to provide a powerful segmentation model [38]. First of all, the input image is splitted into fixed-size patches, linearly embedded each of them, added position embeddings, and feeded the resulting sequence of vectors to a standard Transformer encoder. Concretely, the Transformer, accepts a 1D sequence of feature embeddings as input. There are two variants of the encoder "T-Base" and "T-Larg"" with 12 and 24 layers respectively.

3.1.2 Contrastive representation learning networks

Few attempts have been made in literature to approach semantic segmentation under the fully unsupervised setting. Van Gansbeke et al. [34] applied to a large-scale dataset (i.e. PascalVOC [8]), which is a dataset of roughly 10k images and it aims to decouple feature learning from clustering. Following the same approach, training on the network iMaterialist Fashion Attribute dataset has been performed. The first algorithm proposed a method named MaskContrast [34] which consists of two steps. First, a precedent was defined by identifying objects in the images for which pixels can be grouped together. Mid-level visual groups, like objects, transfer well across datasets, since they do not depend on any pre-defined ground-truth classes. Then, the obtained prior was employed in a contrastive loss [12, 14] to generate pixel embeddings. More specifically, pixels belonging to the same object are pulled together and contrast them against pixels from other objects. The second model explored involves pixel-wise contrastive learning based on semantic segmentation (SemanticSeg) [35]. The main idea behind this method is that the current segmentation models learn to map pixels to an embedding space, ignoring intrinsic structures of labeled data (i.e., inter-image relations among pixels from a same class). Pixel-wise contrastive learning is introduced to foster a new training paradigm, by explicitly addressing intra-class compactness and inter-class dispersion. Each pixel (embedding) i is pulled closer to pixels of the same class but pushed

far from pixels from other class. Thus a better-structured embedding space (e) is derived, eventually boosting the performance of segmentation models.

3.1.3 Fashion image dataset

The dataset chosen for the experiments is the iMaterialist (Fashion) 2020, used for the competition: iMaterialist Challenge (Fashion) at FCVC7 2020[1]. This dataset contains images of people wearing a variety of clothing types in a variety of poses, in daily-life, celebrity events, and online shopping. It consists of approximately 40,000 images and corresponding fashion/apparel segmentation. This dataset contains 46 apparel objects (27 main apparel items and 19 apparel parts) and 294 related fine-grained attributes; for the purpose of this work, only the main apparel items were considered as classes and the fine-grained attributes were not used. The dataset was split into 70% for the training set, 20% for the validation set and 10% for the test set.

3.2 Recommendation module

The core of the recommendation module is based on a similarity-based matching algorithm that uses feature vectors extracted from fashion items. Once the Background Removal Module processes an image, extracting and isolating the fashion item from its background, the Recommender Module uses a deep learning model to analyze the item's visual features, including color, texture, pattern and shape. This model is a Convolutional Neural Network (CNN) fine-tuned for the specific task of fashion item recognition and similarity assessment.

Recognizing a product from its image can be formulated as an image retrieval task, which involves finding images from a database that are most similar to a given query image. Image retrieval can be conceptualized as a problem of measuring vector similarity within a high-dimensional space of image features. CNNs, trained for image classification tasks, serve as effective feature extractors for this purpose. These networks can learn representations of input data during training, mapping input images to real number vectors known as embedding vectors. These embedding vectors aim to encapsulate the semantic or syntactic properties of objects, with similar objects being associated with similar vectors.

In this study, different state-of-the-art classification CNNs were used: VGG16 [31], ResNet [15], Inception [33] and MobileNet [17], which were pre-trained on the ImageNet [7] dataset. The final classification layer of these models was replaced by a dropout layer followed by a dense layer of size 256. Consequently, the output of the models consisted of 256-dimensional vectors, which were further subjected to L2 normalization to constrain values to the range 0 to 1 for ease of comparison. The training used the Triplet Hard Loss [29] function with a soft bound of 1.0, which is particularly effective for learning representations by comparing distances between data points. This loss function operates on triplets of images (anchor, positive, negative) and aims to minimize the distance between the anchor and the positive image (similar products) and maximize the distance between the anchor and the negative image (dissimilar products). The 'hard' in 'triplet hard loss' refers to the selection strategy for triplets [16], focusing on those where the negative image is closer to the anchor than the positive image in feature space. Such triplets are considered to be the most informative for training, as they challenge the model where it is prone to error. This approach, known as hard negative mining, improves the model's ability to discriminate between similar-looking

[1] https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/data

but different classes, thus improving overall performance. Online triplet mining, as proposed by [16], was employed in this study. This approach significantly improves model accuracy and reduces training time by dynamically selecting informative triplets during training. After model training, a testing phase was performed to evaluate the quality of the learned features for product recognition. For this purpose, the test set was used, where embedding vectors were extracted image by image and compared with a gallery of reference images. Cosine similarity was chosen as the criterion for assessing the similarity between the embedding vectors of the query images and those in the gallery.

$$cosine\,similarity = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{1}$$

Cosine similarity quantifies the degree of alignment between two vectors and provides a score between -1 and +1, with values closer to +1 indicating greater similarity. The gallery element with the highest cosine similarity to the query image was designated as a candidate match.

We chose VGG-16 as the architecture. The choice of VGG was made after a quantitative comparison with the other networks (as detailed in Section 4.3). Its effectiveness is motivated by its simplicity, depth and effectiveness in learning complex features from images, making it well-suited to the nuanced task of classifying and matching fashion items.

3.2.1 E-tailer dataset

Our data extraction process involved an advanced web scraping methodology executed in a Linux environment designed to handle the complexities of web architectures. To ensure a comprehensive dataset representative of current fashion trends and offerings, we targeted well-known luxury e-tailers known for their curated selections and high-quality products. This approach used Python, with a focus on the Selenium library, to automate browser interactions. These interactions emulated human navigation of fashion websites, a process critical for accessing dynamic content often rendered via JavaScript. In addition, the BeautifulSoup library played a key role in our methodology, leveraging its ability to analyze HTML and XML content. It transformed the raw, often unstructured data scraped from the web into a more coherent and structured format, making it suitable for detailed analysis. This parsing and cleansing process was critical in reducing the data to its most informative and relevant components, ensuring the integrity of the dataset and its usability for training our classification model. In addition, we carefully labeled each image in the dataset according to its corresponding fashion item category, thus enabling supervised learning and facilitating the training of our classification model. The resulting balanced dataset, with its labeled images distributed across different fashion item categories, is summarized in Table 1.

Table 1 E-tailer dataset for fashion item classification

Class	Number of Images
Tops	200
Bottoms	200
Dresses	200
Shoes	200
Bags	200
Total number of images	**1000**

3.3 Ethics and sustainability module

In the development of the Ethics and Sustainability Module within OutfitAI, we carefully consider the ethical aspects outlined in previous research to ensure responsible and sustainable practices throughout our platform [11]. Drawing from the insights provided by the paper (Table 2, which highlights various technical aspects, ethical concerns, and challenges across different domains of AI and machine learning, we integrate key considerations into our module's design and implementation.

In line with concerns about bias in algorithms and over-reliance on data-driven decisions, we prioritize fairness and transparency in our recommendation engine. We implement mechanisms to mitigate algorithmic bias and ensure diverse representation in our training data to promote fair outcomes for users. Given the privacy concerns associated with image data and the challenge of achieving accuracy in different contexts, we use privacy-preserving techniques and robust image processing algorithms. Our system balances efficiency and privacy, prioritizing user confidentiality while providing accurate and relevant fashion recommendations. To address concerns regarding originality and intellectual property issues in design generation, we establish clear guidelines for the ethical use of AI-generated content. We emphasize the importance of human-AI collaboration in the creative process, fostering a culture of innovation while respecting creators' rights and acknowledging the contributions

Table 2 Overview of AI in fashion: technical aspects, ethical concerns, and challenges

Category	Technical aspects	Ethical concerns	Challenges
Machine learning & Deep learning	- Predictive analytics - Image recognition - Consumer behavior analysis	- Bias in algorithms - Over-reliance on data-driven decisions	- Ensuring accuracy - Diversifying training data
	- Image Recognition & Processing	- Privacy concerns with image data - Accuracy in diverse contexts	- Handling diverse fashion styles - Balancing efficiency and privacy
	- Use of CNNs for fashion item identification - Visual search in e-commerce	- Originality and intellectual property issues - Over-dependence on AI for creativity	- Defining authorship - Encouraging human-AI collaboration
NLP for customer interaction	- Chatbots for customer service - Sentiment analysis in customer feedback	- Privacy and data handling - Misinterpretation of complexities in language	- Developing context-aware systems - Protecting consumer data
	- Supply Chain Optimization	- Impact on employment - Ethical sourcing and production	- Balancing automation and human labor - Transparent supply chain practices
Sustainability	- Resource optimization - AI-driven material selection	- Environmental impact of AI operations - Promoting sustainable practices	- Reducing AI's carbon footprint - Implementing eco-friendly solutions

of both humans and algorithms. Consistent with our goal to promote sustainable practices and reduce AI's carbon footprint, our platform incorporates AI-driven material selection and resource optimization techniques. We prioritize environmental sustainability in our operations and actively seek green solutions to minimize the environmental impact of our AI operations.

4 Results and discussions

In this section, we present the results and discuss the implications of the three key modules that make up the OutfitAI system: the Background Removal Module, the Recommendation Module, and the Ethics and Sustainability Module. Each module plays a critical role in enhancing the functionality and effectiveness of the system, contributing to its overall utility and impact in the fashion domain. While the proposed deep learning models, such as transformer-based architectures and contrastive learning techniques, offer significant advantages in terms of accuracy, robustness and generalisation, they are not without limitations. These models can have high computational and memory requirements, leading to potential latency issues, especially in real-time applications or when deployed in resource-constrained environments. Transformer models require significant memory due to their attention mechanisms, and contrastive learning methods can introduce computational overhead during training due to the need for large batches of data. Despite these challenges, the performance advantages of these models, such as their ability to handle complex tasks like semantic segmentation with high accuracy and robustness, overcome the drawbacks. The strong generalisation capabilities and robust feature representation make these models invaluable for applications where accuracy and data complexity are paramount.

First, we analyse the performance of the Background Removal Module, which uses advanced deep learning techniques, including unsupervised semantic segmentation and transformer-based neural networks, to seamlessly remove backgrounds from fashion images. We evaluate the accuracy and efficiency of each approach and discuss their suitability for real-world applications, considering factors such as computational resources and scalability. We then discuss the recommendation module. We present the results of our recommendation engine's performance, including metrics such as precision, recall and F1-score. We also explore the challenges and opportunities associated with personalized fashion recommendations, considering factors such as user diversity and evolving fashion trends. Finally, we examine the ethics and sustainability module, which incorporates ethical considerations and sustainability principles into the design and operation of the OutfitAI" platform. We discuss the ethical implications of AI-driven fashion recommendations and the strategies employed to promote fairness, transparency and environmental responsibility.

4.1 Background removal module results

4.1.1 Transformers settings

During training, the standard pipeline from the semantic segmentation library MMSegmentation was followed, as it provides mean subtraction, random resizing of the image to a ratio between 0.5 and 2.0 and random left-right flipping. Then, a random crop has been applied on large images and a padding has been applied on small images, to make them suitable for a fixed size of 512×512, which was the size used by models pre-trained on ADE20K dataset.

Each architecture has done a distributed training of 15 epochs across 2 RTX 2080 GPUs. The batch size for Segmenter ViT-T and ViT-B has been set to 8 and 4 respectively. Instead, the batch size of SETR-PUP and MLA has been set to 2 for both. As transformer-based models have a higher memory occupation, it was not possible to choose a larger batch size. A Vision Transformer encoder accepts fixed-size patches as input images, hence the size of these patches has been set to 16. To fine-tune the pre-trained models for the semantic segmentation task, the standard pixel-wise cross-entropy loss was used, without weight re-balancing. Furthermore, the stochastic gradient descent (SGD) was employed as optimizer, with a polynomial learning rate decay, starting with a learning rate set to 0.001 and a momentum of 0.9 for all architectures.

4.1.2 Contrastive representation learning settings

Firstly, some experiments have been conducted to test the correct configuration of the algorithm with the original datasets; then, the fine-tune on the iMaterialist dataset was done. Concerning MaskContrast, the dataset used for initial configuration purposes is PascalVOC [8] which contains a total of 17.125 images divided into 20 classes plus one for background. The total images are splitted into 10.582 images belonging to the training set and 1.449 images belonging to the validation set. Regarding the ContrastiveSeg model, the original dataset which it was trained on is called Cityscape [6] and is comprised of 5.000 finely annotated urban scene images dived into 2.975 training set, 500 validation set and 1.524 testing set. This dataset was used for initial configuration purposes. As noted before, the dataset used to conduct the experiments for both networks is iMaterialist Fashion Attribute dataset, which provided a large number of images and corresponding fashion/apparel segmentations.

In the original ContrastiveSeg training, various backbones (i.e., ResNet and HRNet) and segmentation networks (i.e., DeepLabV3, HRNet, and OCR) are exploited in the experiments to validate the proposed algorithm. Conventions are followed for training hyper-parameters. All the backbones are initialized using corresponding weights pretrained on ImageNet [28] with the remaining layers being randomly initialized. For data augmentation, the algorithm implements color jitter, horizontal flipping and random scaling with a factor in [0.5, 2]. As optimizer, SGD has been used, with a momentum of 0.9 and weight decay of 0.0005. For the experiments concerning this work, first of all, the ground truth of the images have been generated. Originally, all training images were augmented by random cropping the dimension from 1024×2048 to 512×1024, but, differently from the original dataset, in this case images did not present the same size and were, in fact, larger, in terms of both height and width. Thus, a resize of all the image to 1024×512 has been performed, with a preemptive check on the orientation of the images (i.e., portrait or landscape). Once the dataset was in the correct format training was performed, running for 1000 iterations.

As for the MaskContrast network, first of all, ground truth for the images have been generated using a script; afterwards, starting from these ground truth images, saliency masks have been created using BASNet [25], which consists of a predict-refine network and a hybrid loss, for highly accurate image segmentation; it consists of a U-Net-like [27] deeply supervised, an "heavy" encoder-decoder network and a residual refinement module with "light" encoder-decoder structure. The "heavy" encoder-decoder network transfers the input image to a segmentation probability map, while the "light" refinement module refines the predicted map by learning the residuals between the coarse map and ground truth.

Testing on this network has been divided into four steps:

– Pre-train
– Linear classifier
– Clustering (K-means)
– Semantic Segment Retrieval

The pre-training was done by transferring the information from the Imagenet weights, and fine-tuning them with the iMaterialist dataset. SGD as been used as optimizer with a momentum of 0.9 and weight decay of 0.0001. The learning rate (LR) was set to 0.004 and, to optimize this value, a polynomial scheduler was used. The pre-train ran for 15 epochs.

In the linear classifier step, the weights of the pre-trained model were frozen and a 1 x 1 convolutional layer was trained to predict the class assignments from the generated feature representations. Since the discriminative power of a linear classifier is low, the pixel embeddings need to be informative of the semantic class to solve the task in this way. With the clustering method, a verification on whether the feature representations can be directly clustered in semantically meaningful groups can been performed, using an off-line clustering criterion (e.g., K-Means). The number of clusters equals the number of ground truth classes. The Hungarian matching algorithm is used to match the predicted clusters with the ground-truth classes and the results are averaged across five runs, one round for train, one round for validation. The pre-trained model was given as input.

For the Semantic Segment Retrieval, firstly, a feature vector for every salient object has been computed, by averaging the pixel embeddings within the predicted mask. Then, the nearest neighbors of the validation set objects from the training set have been retrieved. The pre-trained model was give as input. In addition, it was specified to ignore the background class, since it is not needed for this task.

4.2 Deep learning models for background removal results

Table 3 shows quantitative results in terms of accuracy and mIoU for all the models trained.

The models Segmenter ViT-B, SETR-MLA and SETR-PUP have achieved excellent results, in particular Segmenter ViT-B and SETR-MLA. Segmenter ViT-T model, despite showing good accuracy, presents the lowest value of the 4 transformer-based networks (and showed high loss values during training); it can be assumed that, as this model is simpler that the others, it fails to extract the required features for prediction correctly in some particular cases. This shows, as expected, that the encoder based on ViT-L used in SETR models is more performing than the encoder based on ViT-T and ViT-B used in Segmenter models.

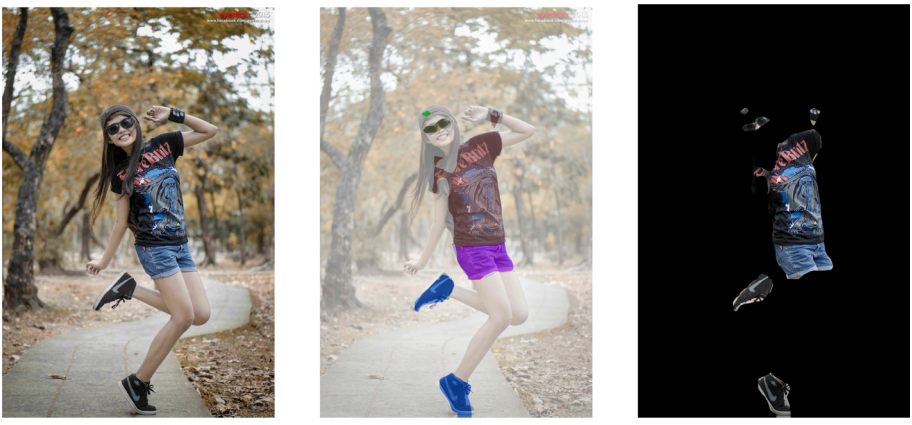
Table 3 Quantitative results comparison between models

	Model	Encoder/Decoder	Accuracy	mIoU
Transformer	Segmenter	ViT-T/Mask Transformer	92.35	38.64
		ViT-B/Mask Transformer	94.76	47.11
	SETR	ViT-L/MLA	94.42	42.59
		ViT-L/PUP	94.22	41.23
Contrastive Representation Learning	MaskContrast	n/a	95.86	20.22
	ContrastiveSeg	n/a	84.8	24.1

Both metrics used confirm that the best architecture is Segmenter ViT-B. As it can seen from the qualitative results of Segmenter ViT-B (that has the best mIoU compared to the others) shown in Fig. 2, most of the main classes (e.g., shorts, t-shirts) can be correctly identified, while the classes of small object such as clock and hat are more challenging for the network. This behavior has a big impact when the mean IoU is calculated, resulting in a score reduction, while the Accuracy is not affected.

Regarding the contrastive representation learning networks, MaskContrast has achieved really good results in terms of accuracy. As with transformer-based architectures, these kind of models are very heavy in terms of size and require a lot of computational power for the training phase. The results achieved during the testing phase are reported in Table 3. Both networks show promising results with the bigger classes e.g., "t-shirt", "sweatshirt", "dress" or "jacket", while, as expected, the networks performed less better with smaller classes e.g., "glasses", "glove" and "scarf" (since these classes are also less represented in the dataset). In particular, for MaskContrast, this behavior can be attributed to the fact that a resize of a factor of 1/7 of the images contained inside the dataset has been performed, in order to avoid overflow on GPUs, with the reduction of details that follows. Also, this behavior impacts the evaluation of the mean IoU. Overall, MaskContrast achieved a better Accuracy result, but a lower mIoU than ContrastiveSeg.

According to the results obtained, the Transformer-based techniques achieved slightly lower accuracy compared to the MaskContrast network. However, the transformer-based techniques greatly outperformed the contrastive representation learning techniques in terms of mIoU. The fact that the transformer-based networks achieved better mIoU suggests that they were able to produce more accurate segmentation masks, despite having slightly lower accuracy than the MaskContrast model. On the other hand, MaskContrast achieved higher accuracy, which indicates that it was better at correctly identifying the objects of interest in the image. However, it produced lower mIoU, indicating that the predicted segmentation masks were not as accurate as those produced by the transformer-based techniques. Overall, the results suggest that both approaches have their strengths and weaknesses, and the choice between them may depend on the specific requirements of the application. If accurate segmentation masks are a priority, then the transformer-based techniques may be more



(a) Original image (b) Segmentation map (c) Background removed

Fig. 2 Qualitative results of segmenter ViT-B

suitable; however, if accurately identifying the object of interest is more important, then the MaskContrast technique may be a better choice.

4.3 Recommendation module

We present a comparison of different CNN networks used in the recommendation module of OutfitAI to evaluate their performance in providing personalized fashion recommendations. The evaluation was performed on the E-tailer dataset collected from luxury e-commerce websites. We evaluated four commonly used CNN architectures: VGG16, ResNet, Inception and MobileNet. Each network was trained using the same dataset and supervised learning approach to develop recommendation engines tailored to user preferences and interests. Performance metrics including precision, recall and F1 score were calculated to assess the effectiveness of each CNN network in recommending relevant fashion items to users. Precision measures the proportion of recommended items that are relevant to the user's preferences, while recall quantifies the proportion of relevant items successfully retrieved by the system. The F1 score provides a balanced assessment of precision and recall, taking into account both the model's ability to recommend relevant items and its efficiency in filtering out irrelevant items. The comparison results, summarized in Table 4, show that VGG16 outperformed other CNN networks in terms of precision, recall and F1 score. VGG16 achieved a precision of 0.85, a recall of 0.80 and an F1 score of 0.82, demonstrating its superior ability to accurately recommend relevant fashion items to users. ResNet, Inception and MobileNet also performed reasonably well, but fell short of VGG16 on all metrics.

The superior performance of VGG16 highlights its effectiveness in capturing complex features and patterns in fashion images, thereby improving the accuracy and relevance of fashion recommendations. The results underscore the importance of selecting an appropriate CNN architecture tailored to the specific requirements of fashion recommendation tasks. While ResNet, Inception and MobileNet are viable alternatives, their performance may vary depending on factors such as dataset characteristics and model complexity.

4.4 Ethics and sustainability module results

The resulting framework is designed to integrate ethical considerations and sustainability into the core of business practices. Below is a hypothetical framework that could result from such a module, focusing on the principles that participants were encouraged to integrate into their daily work routines and organizational policies. Our ethical framework for the fashion module of OutfitAI is based on four main principles: Transparency and Explainability, Fairness, Reliability and Sustainability. These principles guide our approach to developing and deploying deep learning (DL) systems for fashion recommendation, ensuring our platform operates ethically and responsibly while promoting positive user experiences and outcomes.

Table 4 Comparison of CNN networks for fashion recommendation

CNN network	Precision	Recall	F1-Score
VGG16	0.85	0.80	0.82
ResNet	0.82	0.75	0.78
Inception	0.78	0.72	0.75
MobileNet	0.79	0.74	0.76

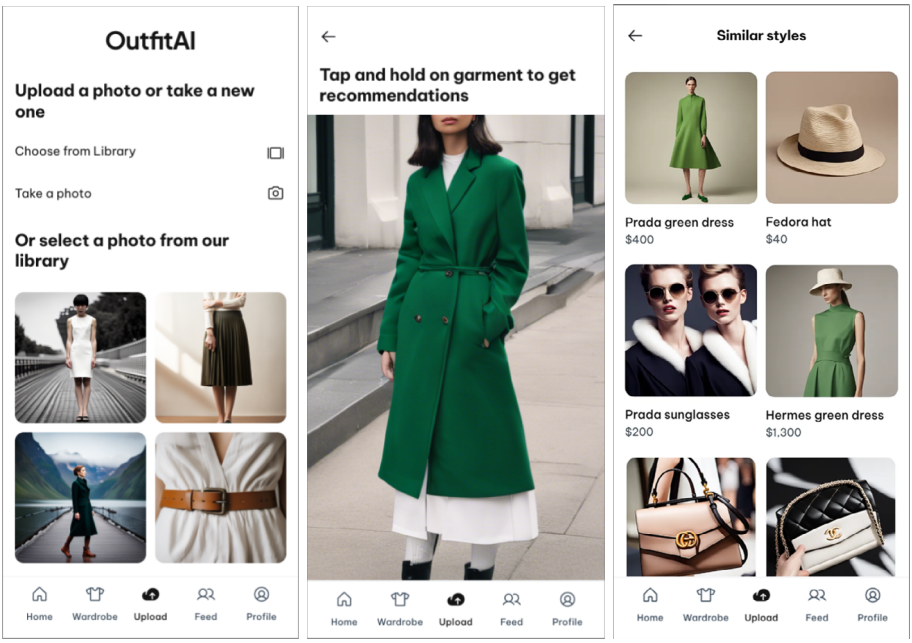
The ethical framework outlined in Table 5 serves as a guiding principle for the development and operation of OutfitAI. Our platform aims to provide users with ethical and responsible fashion recommendations by prioritizing transparency, fairness, reliability and sustainability. Transparency and explainability ensure that users can understand and interpret the recommendations provided by OutfitAI", thus fostering trust in the platform. Fairness is paramount to avoid bias or discrimination in recommendation outcomes and ensure equal access to fashion recommendations for all users. Reliability emphasizes the importance of accurate and reliable data, minimizing the risk of misinformation or misleading recommendations. Sustainability promotes eco-conscious fashion choices, aligning with ethical sourcing and environmentally friendly practices. Together, these principles contribute to the trustworthiness of OutfitAI", providing users with reliable, ethical and sustainable fashion recommendations that empower them to make informed and responsible choices in their fashion journey.

5 OutfitAI platform

In the technical design of the visualization platform for OutfitAI, we have carefully built a robust and scalable architecture using a combination of the latest tools and technologies. The frontend is developed using modern web frameworks such as React.js and styled using CSS preprocessors such as Sass, ensuring a responsive and intuitive interface for users to explore fashion recommendations. On the backend, we use Node.js as the runtime environment and Express.js as the web application framework to handle server-side logic and API integrations. MongoDB serves as the database management system, providing a flexible and scalable solution for storing and managing user data. In addition, image processing is performed using libraries such as OpenCV and Pillow to enhance the visual presentation of fashion items. Rigorous testing and quality assurance processes are carried out using tools such as Jest, Mocha and Selenium to validate the functionality and reliability of the platform. To help users navigate the platform effectively, a comprehensive user manual is provided with step-by-step instructions, helpful tips and best practices to maximize the platform's features and capabilities. Through the strategic use of these tools and technologies, OutfitAI empowers

Table 5 Ethical Framework for the Fashion Module of "OutfitAI"

Principle	Description
Transparency and explainability	Prioritize transparency and explainability in AI models and recommendation processes to ensure understandability and interpretability of recommendations.
Fairness	Ensure fairness in recommendation outcomes, avoiding bias or discrimination based on demographic factors such as gender, race, body type, and socioeconomic status.
Reliability	Emphasize the accuracy and reliability of data used to train recommendation models, minimizing the risk of misinformation or misleading recommendations.
Sustainability	Align DL techniques with sustainable fashion practices, promoting ethically sourced and environmentally friendly fashion products.
Trustworthiness	Ensure technical and social robustness, providing reliable and ethical fashion recommendations that integrate seamlessly into users' lives.



(a) Home page, where a user can upload a picture to be injected in the framework

(b) Item selection page, where the picture taken is shown and a user can highlight a garment to get a recommendation based on that item

(c) Recommendation page, where results are shown, with products having name and price attached, based on online shop.

Fig. 3 OutfitAI platform

users to make informed and sustainable fashion choices, enhancing their overall shopping experience.

Figure 3 illustrates the architecture of the platform for visualizing results in OutfitAI.

6 Conclusions and future works

This paper have proposed a comprehensive exploration of OutfitAI, a sophisticated system designed to revolutionize the fashion industry by leveraging deep learning techniques, ethical considerations, and user-centric design principles. Through careful research and development, we have developed a robust system architecture consisting of three main modules: the Background Removal Module, the Recommendation Module, and the Ethics and Sustainability Module. The Background Removal Module uses state-of-the-art deep learning techniques, including unsupervised semantic segmentation and transformer-based neural networks, to accurately identify and remove backgrounds from fashion images. This module plays a critical role in enhancing the visual presentation of fashion items and providing personalized recommendations to users. The Recommendation Module harnesses the power of CNNs, such as VGG16, to analyze fashion images and make personalized recommendations based on the user's preferences and style. The ethics and sustainability module outlines a

comprehensive ethical framework for the development and operation of OutfitAI, addressing key ethical risks associated with deep learning systems in the fashion domain. By prioritizing transparency, fairness, trustworthiness and sustainability, the module ensures that the platform operates ethically and responsibly, fostering trust among users.

Looking ahead, there are several opportunities for future research and development to further enhance the capabilities and impact of OutfitAI. Firstly, ongoing advances in deep learning algorithms and image processing techniques can improve the accuracy and efficiency of the background removal and fashion recommendation processes. In addition, expanding the platform's integration with external datasets and APIs can provide users with more comprehensive insights into fashion trends, sustainability metrics and brand authenticity. Furthermore, ongoing efforts in user experience (UX) design and usability testing can refine the platform's interface and functionality to ensure a seamless and intuitive experience for users across different demographics and preferences. Incorporating user feedback and iterative improvements can drive continuous innovation and refinement of the platform, increasing its value proposition and user engagement over time. To address the limitations discussed, future work will also focus on exploring optimisation techniques such as model pruning and quantization. These methods have the potential to significantly reduce the memory and computational cost of deep learning models without sacrificing performance. Model pruning can help eliminate redundant weights and parameters, leading to a more efficient model, while quantization can reduce the precision of weights, allowing for faster inference and lower memory consumption. By integrating these optimisation strategies, we aim to make models more suitable for real-time applications and environments with limited computational resources. In addition, we will explore methods to improve the training efficiency of contrastive learning models, possibly through advanced augmentation techniques and more efficient contrastive loss functions. These improvements will ensure that the benefits of the models continue to outweigh their limitations, making them both highly effective and resource efficient in practical applications.

Acknowledgements This research was funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MIUR), National Innovation Ecosystem grant ECS00000041-VITALITY-CUP D83C22000710005.

Funding Open access funding provided by Università Politecnica delle Marche within the CRUI-CARE Agreement.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Amanlou A, Suratgar AA, Tavoosi J, Mohammadzadeh A, Mosavi A (2022) Single-image reflection removal using deep learning: a systematic review. IEEE Access
2. Annunziata E, Pucci T, Cammeo J, Zanni L, Frey M (2023) The mediating role of exogenous shocks in green purchase intention: evidence from italian fashion industry in the covid-19 era. Italian Journal of Marketing pp. 1–21
3. Balim C, Özkan K (2023) Diagnosing fashion outfit compatibility with deep learning techniques. Expert Syst Appl 215:119305
4. Bouwmans T, Javed S, Sultana M, Jung SK (2019) Deep neural network concepts for background subtraction: a systematic review and comparative evaluation. Neural Netw 117:8–66
5. Bouwmans T, Silva C, Marghes C, Zitouni MS, Bhaskar H, Frelicot C (2018) On the role and the importance of features for background modeling and foreground detection. Computer Science Review 28:26–91
6. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223
7. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
8. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vision 88:303–338
9. Fang W, Ding Y, Zhang F, Sheng VS (2019) Dog: a new background removal for object recognition from images. Neurocomputing 361:85–91
10. Garcia-Garcia B, Bouwmans T, Silva AJR (2020) Background subtraction in real applications: challenges, current models and future directions. Computer Science Review 35:100204
11. Giovanola B, Tiribelli S, Frontoni E, Paolanti M (2023) Ethical implications of artificial intelligence in the fashion industry: a comprehensive analysis. Fashion Highlight 2:22–28
12. Gutmann M, Hyvärinen A (2010) Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 297–304. JMLR Workshop and Conference Proceedings
13. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y et al (2022) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87–110
14. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
16. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
17. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
18. Jain NK, Saini R, Mittal P (2019) A review on traffic monitoring system techniques. Soft computing: theories and applications: Proceedings of SoCTA 2017:569–577
19. Kang MS, An YK (2021) Deep learning-based automated background removal for structural exterior image stitching. Appl Sci 11(8):3339
20. Kinli F, Ozcan B, Kirac F (2021) Instagram filter removal on fashionable images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 736–745
21. Liang J (2022) The research of background removal applied to fashion data: the necessity analysis of background removal for fashion data
22. Liang J, Liu Y, Vlassov V (2023) The impact of background removal on performance of neural networks for fashion image classification and segmentation. arXiv:2308.09764
23. Miyamoto R, Nakajima T, Oki T (2019) Accurate fashion style estimation with a novel training set and removal of unnecessary pixels. In: 2019 IEEE international symposium on circuits and systems (ISCAS), pp 1–5. IEEE
24. Oyewusi WF, Onilude G, Adekanmbi O, Akinsande O (2020) Afrirazer: a deep learning model to remove background and skin from traditional african fashion images. In: Paper presented at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

25. Qin X, Fan DP, Huang C, Diagne C, Zhang Z, Sant'Anna AC, Suarez A, Jagersand M, Shao L (2021) Boundary-aware segmentation network for mobile and web applications. arXiv:2101.04704
26. Ramé A, Douillard A, Ollion C (2022) Core: color regression for multicolor fashion garments. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2252–2257
27. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp 234–241. Springer
28. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252
29. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
30. Shimizu R, Saito Y, Matsutani M, Goto M (2023) Fashion intelligence system: an outfit interpretation utilizing images and rich abstract tags. Expert Syst Appl 213:119167
31. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
32. Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7262–7272
33. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9
34. Van Gansbeke W, Vandenhende S, Georgoulis S, Van Gool L (2021) Unsupervised semantic segmentation by contrasting object mask proposals. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10052–10062
35. Wang W, Zhou T, Yu F, Dai J, Konukoglu E, Van Gool L (2021) Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7303–7313
36. Yu F, Wang D, Chen Y, Karianakis N, Shen T, Yu P, Lymberopoulos D, Lu S, Shi W, Chen X (2022) Sc-uda: style and content gaps aware unsupervised domain adaptation for object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 382–391
37. Zhang J, Fukuda T, Yabuki N (2021) Automatic object removal with obstructed façades completion using semantic segmentation and generative adversarial inpainting. IEEE Access 9:117486–117495
38. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PH et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6881–6890
39. Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, Torralba A (2019) Semantic understanding of scenes through the ade20k dataset. Int J Comput Vision 127:302–321