IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 July 2023, accepted 11 August 2023, date of publication 17 August 2023, date of current version 23 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3306235

RESEARCH ARTICLE

AI Assisted Fashion Design: A Review

ZIYUE GUO[1], ZONGYANG ZHU[1], YIZHI LI[1], SHIDONG CAO[1], HANGYUE CHEN[2], AND GAOANG WANG[1], (Member, IEEE)

[1]Zhejiang University-University of Illinois Urbana-Champaign Institute (ZJUI), Zhejiang University, Haining 264199, China
[2]Hangzhou Dianzi University, Hangzhou 310005, China

Corresponding authors: Hangyue Chen (chy@hdu.edu.cn) and Gaoang Wang (gaoangwang@intl.zju.edu.cn)

This work was supported by the Natural Science Foundation of Zhejiang Province under Grant QY19E050003.

ABSTRACT This review explores the integration of enhanced personalization and seamless multimodal interfaces in the field of fashion design and recommendation. We examine the increasing demand for personalized fashion experiences and the potential of multimodal interfaces in facilitating effective communication between designers and users. By leveraging user preferences, body measurements, and style choices, artificial intelligence (AI) systems can deliver highly personalized fashion recommendations. The integration of various input modalities, including text, images and sketches, enables designers and users to communicate their design ideas with ease. The primary results highlight the transformative potential of enhanced personalization and seamless multimodal interfaces, empowering designers and consumers to co-create unique and personalized designs. This paradigm shift fosters a deeper level of engagement and creativity within the fashion industry. Embracing this advancement unlocks unprecedented opportunities for designers, brands, and consumers, ushering in a new era of innovation and creativity in fashion design.

INDEX TERMS Artificial intelligence, deep learning, fashion design.

I. INTRODUCTION

Fashion is a dynamic and influential form of self-expression and cultural representation, shaped by historical events, cultural movements, media, and technology. It plays a significant role in society, impacting personal identity, social interactions, and the economy. Meanwhile, artificial intelligence (AI) has garnered considerable attention, particularly in image processing [1], [2], [3], [4], [5], with notable advancements in deep learning and generative models driven by the prevalence of images in social media [4], [6]. Nowadays, AI and fashion design have developed a strong and evolving relationship. AI technologies are being increasingly utilized in the fashion industry to enhance various aspects of the design process, from fashion detection, synthesis to recommendation. It can empower fashion designers with tools and insights that streamline the design process, enhance creativity, and meet the evolving demands of consumers. It serves as a valuable ally in the fashion industry, driving innovation, efficiency, and sustainability. The main objective of this paper is to introduce the application and development of AI in the field of fashion design.

Our review focuses on the application of artificial intelligence in the field of fashion design, summarizing more than sixty recent research achievements at the intersection of fashion and computer vision. The review encompasses a comprehensive overview, ranging from traditional methods to deep learning techniques, revealing the diversity and innovation of AI technologies within the fashion domain. It includes literature of various types, including journal articles, conference papers, and preprints. The covered literature spans from the year 2011 to 2023, showcasing the recent research trends in this field.

We divide the field of fashion design into three major parts: fashion detection [7], [8], [9], [10], fashion synthesis [2], [11], [12], [13], and fashion recommendation [4], [14]. In addition, we also give an overview of main applications in these fashion domains, showing the strengths and shortcomings of intelligent fashion in the fashion industry. The reason why we organize the survey in this structure is to simulate the working process of a fashion designer. In most design work, designers first analyze fashion trends, gather inspiration, and

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

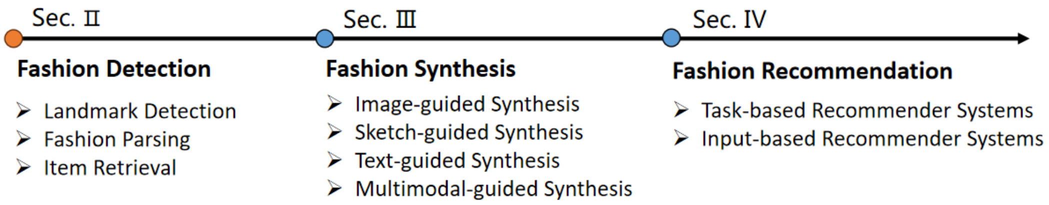


FIGURE 1. **Taxonomy of the survey, including fashion detection, synthesis and recommendation.**

extract and categorize fashion elements. Then they proceed to the fashion synthesis process, where they design new fashion items or modify existing ones. Once the fashion items are created, designers also need to provide recommendations for outfit combinations based on user preferences and usage scenarios.

Before delving further into the intricacies of our survey, it is essential to review the relevant literature and consolidate the existing knowledge on this topic. Gu et al. [15] proposed a survey focused on fashion analysis and understanding with artificial intelligence from 2011 to 2019. Nonetheless, a fresh and comprehensive survey is yet to emerge to consolidate contemporary methodologies, appraise evaluation metrics, and offer valuable insights into prospective avenues of research. Moreover, we have focused our selection of research papers to concentrate on specific areas. Focusing on augmenting the introductory accomplishments, our survey even incorporates advancements up to the year 2023, highlighting significant advancements such as multimodal approaches [13], large-scale models, and diffusion models [3] for AI.

In addition, we are inspired by the survey proposed by Cheng et al. [14], especially the classification method. The survey introduced more than 200 major fashion-related works published from 2012 to 2020. However, our review focuses more on the field of fashion design, so we have refined the classification and supplemented the new developments in the field on a large scale. Mohammadi et al. [16] extensively studied AI uses in fashion and apparel, analyzing over 580 articles across 22 fashion tasks using a structured multi-label classification approach. The major characteristic of the survey lies in the vast number of categories, while our comprehensive survey offers a higher level of summarization and greater concentration of content.

Overall, the contributions of our work can be summarized as follows:

- We present a comprehensive survey that examines the current research progress in the field of fashion design. We categorize the research topics into three main categories: detection, synthesis, and recommendation.
- For each category, we conduct an in-depth and organized review of the most significant methods and their respective contributions and existing limitations and shortcomings.
- Lastly, we outline existing challenges together with the possible future directions that can contribute to further advancements in the field.

Section II reviews the tasks of fashion detection, which include landmark detection, fashion parsing, and item retrieval. Section III provides an overview of fashion synthesis, which focuses on assisting designers in creating and improving fashion items. Section IV discusses the works of fashion recommendation, specifically on fashion items matching and innovative recommender systems.

II. FASHION DETECTION

Fashion detection [7] is the automated analysis and recognition of various fashion-related attributes, elements, and categories within images. The primary objective of fashion detection in clothing is to enable automated analysis, understanding, and interpretation of fashion-related elements, contributing to a range of applications.

The improvement of the detection model in fashion design plays an important role in enabling the efficient analysis of extensive datasets comprising fashion images. By accurately detecting and classifying various fashion elements, including clothing items, styles, and patterns, the reliance on manual identification by designers is alleviated. This improvement not only enhances productivity but also grants designers the freedom to devote more time to creative pursuits such as ideation and design. Hence, the pursuit of fashion detection improvement bears significant importance within the industry. However, traditional detection methods [17], [18], [19], developed over several decades, exhibit certain limitations. These methods typically involve a three-step process encompassing region suggestion, feature extraction, and classification and regression. Yet, they are plagued by computationally intensive calculations during the region suggestion stage and are confined in their ability to capture nuanced, high-level features during feature extraction. Moreover, the process's division into distinct stages renders the search for a global optimal solution unviable. Consequently, substantial improvements in fashion detection are imperative to propel advancements in the field of fashion design, addressing the aforementioned shortcomings and opening avenues for more efficient and accurate analysis of fashion imagery.

We divided this section into three subsections: landmark detection, fashion parsing and item retrieval. Fashion detection is to accurately identify and localize landmarks on a user's body for precise mapping onto virtual clothing items. Fashion parsing is the process of analyzing and segmenting images to extract fine-grained information about clothing. And item retrieval combines historical data, visual content, and fashion attributes to provide personalized and

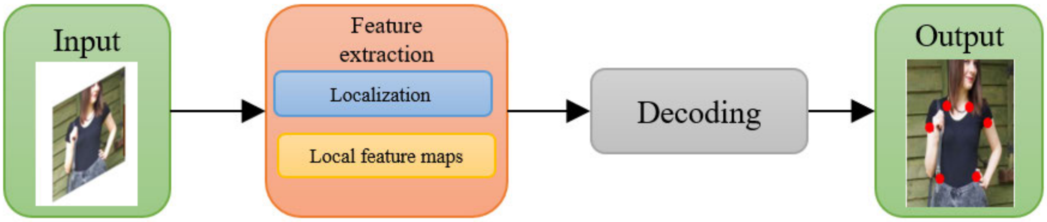


FIGURE 2. Illustration of landmark detection with input and output.

accurate recommendations. For each subsection, we will first introduce their concept and processing, and then share some improvement within the past few years.

A. LANDMARK DETECTION

1) OVERVIEW

Landmark detection [8] is a crucial component in fashion detection, serving the purpose of accurately identifying and localizing landmarks on a user's body for precise mapping onto virtual clothing items. This process plays a pivotal role in generating realistic depictions of how garments would fit and appear on individuals, especially in virtual fitting rooms. By comprehending body shape and proportions through landmarks such as shoulders and hips, landmark detection significantly enhances the virtual try-on experience. It facilitates the seamless integration of virtual garments with the user's body, resulting in immersive virtual shopping interactions and advancing the field of fashion design. Figure 2 is a simple illustration of landmark detection process. The visibility of landmarks would be determined firstly. And through the convolutional neural networks (CNN) model, the local feature maps around the landmark location could be obtained. Then they were aggregated to produce the final feature maps. The final feature maps then went to decoding, and we can estimate the classes and attributes of clothes.

2) DEVELOPMENT

Landmark detection emerged as a concept proposed by Liu et al. [8] to predict crucial positions of fashion items, such as necklines corners, hemlines, and cuffs, and facilitate fashion clothing retrieval. This approach has found extensive usage within the fashion industry, effectively addressing challenges associated with capturing clothing features. Researchers have undertaken significant endeavors to enhance the applicability of landmark detection in fashion analysis. Wang et al. [20] introduced a fashion grammar model that combines the learning capabilities of neural networks with domain-specific grammars, aiming to tackle issues related to fashion landmark localization and clothing category classification. This model successfully captures kinematic and symmetric relationships between clothing landmarks, yielding improved accuracy in landmark localization.

In addition to the fashion grammar model, researchers have proposed innovative methods to further refine landmark detection in fashion analysis. Huang et al. [21] presented a novel deep end-to-end architecture based on part affinity fields (PAFs) for landmark localization. This method employs a stack of convolution and deconvolution layers to generate initial probabilistic maps of landmark locations, which are subsequently refined through the exploitation of associations between landmark locations and orientations. Notably, this approach yields notable enhancements in clothing category and attribute prediction. Moreover, Lee et al. [22] introduced a global-local embedding module into landmark detection, effectively predicting landmark heatmaps and leveraging comprehensive contextual knowledge of clothing. This module adeptly handles challenges posed by substantial variations and non-rigid deformations in clothing images, significantly improving the accuracy of landmark detection.

These advancements in landmark detection techniques have made significant contributions to the field of fashion analysis, facilitating precise and reliable identification of fashion landmarks within images. Through the integration of neural networks, domain-specific grammars, and the exploitation of contextual knowledge, researchers have achieved remarkable progress in enhancing the accuracy and efficacy of fashion landmark detection. Ultimately, these advancements deepen the detection and analysis of fashion items.

B. FASHION PARSING

1) OVERVIEW

Fashion parsing is the computational procedure of analyzing and segmenting images or videos in order to extract fine-grained information pertaining to clothing and fashion-related elements. This intricate process entails the identification and categorization of diverse fashion attributes, encompassing garment types, collars, necklines, patterns, and textures, among others. By effectively unraveling the intricacies of fashion representations, parsing facilitates a deeper understanding and organization of fashion-related data, thereby significantly contributing to the advancement of fashion detection and optimization endeavors. The accurate parsing of fashion imagery yields valuable insights that find application in various domains, including personalized styling, fashion recommendation systems, and trend analysis. The procedure of fashion parsing is as shown in figure 3. First the model would use a pre-trained backbone to extract features from the input image. After decoding, the recovered features are then used for the contour prediction of the person in the edge branch and the person segmentation in the parsing branch and we can get the prediction result of parsing.

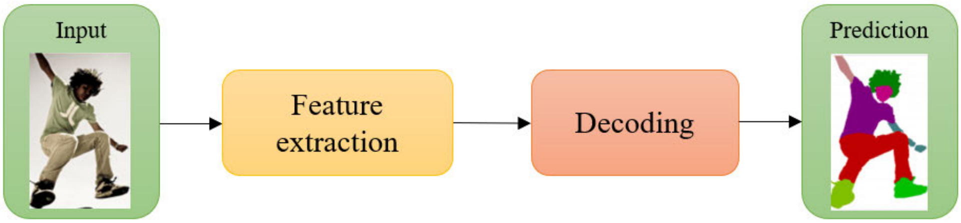


FIGURE 3. Illustration of fashion parsing process with input and prediction.

2) DEVELOPMENT

Since Yamaguchi et al. [9] put forward fashion parsing, it quickly became popular among fashion detection. During the fashion parsing, clothing labels could be predicted according to to body parts. Then Yamaguchi et al. [23] proposed a retrieval-based approach. The process involved retrieving similar images from a parsed dataset based on a given image, and then transferring the nearest-neighbor parsing to the final result using dense matching. However, the initial fashion parsing method only worked under the constrained condition, which means tags should be given to indicate the clothing items. So Yamaguchi et al. [23] also combined the model with pre-trained global models for clothing items, dynamically learning local models from retrieved examples, transferring parse mask predictions from retrieved examples to the query image, and incorporating iterative label smoothing for enhanced output coherence. In this case, labels do not need to be defined in advance, but can be extracted directly from the labels of the pre-trained model, which provides a new idea.

But when there were no or few tags and annotations for the clothing, i.e. under the unconstrained condition, the above approaches could not work well as expected. To make fashion parsing model work better under the unconstrained condition, Liang et al. [24] introduced a joint image segmentation and labeling approach which consists of two phases of inference: image co-segmentation and region co-labeling. Image co-segmentation is for extracting distinguishable clothing regions, and region co-labeling is for recognizing garment items. The approach helps overcome the problem of severe occlusions between clothing items and human bodies, successfully enables precisely locating the region of interest for the query. Liang et al. [25] also combined fashion parsing with CNN model to capture the complex correlations between clothing appearance and structure. Liang et al. [26] then built am improved model called contextualized CNN (Co-CNN), which aimed at simultaneously capture the cross-layer context and global image-level context to improve the accuracy of parsing results.

Thanks to the import of CNN, fashion parsing model was optimized. Ruan et al. [27] used Mask R-CNN [28] to achieve multi-person parsing. And Liu et al. [29] proposed Matching CNN (M-CNN), which solved the issue of human parsing methods depended on the hand-designed pipelines composed of multiple sequential components. Obviously, CNN model greatly helped solved some fashion parsing related problems and it deserved further work.

C. ITEM RETRIEVAL

1) OVERVIEW

In order to cater to the diverse preferences of consumers in clothing purchases, it is necessary to incorporate personalized recommendations based on their past searches and feedback. This integration of historical data has the potential to significantly enhance the efficiency of the search process. Item retrieval is a widely employed technique that allows for the searching and retrieval of visually similar items or images by leveraging their visual content. By combining item retrieval with fashion detection, a more comprehensive approach can be achieved. To recommend appropriate clothing for users, recall and sort play an important role. Recall is the ability of the retrieval system to find and retrieve all relevant fashion items that are similar to the query image. A high recall indicates that the system can successfully identify a significant portion of similar items, ensuring that relevant results are not missed. For example, if a user queries for a specific dress, high recall means the system can find most similar dresses in the database. After that, the system will sort those clothes based on how similar the search results are to the queried image. The items with the highest similarity or relevance are ranked at the top of the list, while less relevant or dissimilar items are ranked lower. To make item retrieval better fit the user's need, some studies explored the incorporation of deep learning methodologies with item retrieval, yielding notable advancements in the effectiveness of personalized recommendations.

2) DEVELOPMENT

The introduction of the item retrieval method has swiftly gained popularity in the recognition of clothing trends, layering techniques, body shape analysis, and postures. Li et al. [10] proposed a two-step approach for cross-scenario retrieval of clothing items and fine-grained clothing style recognition. Firstly, they introduced a hierarchical super-pixel merging algorithm based on semantic segmentation to obtain intact query clothing items. Secondly, to address the challenges of clothing style recognition across different scenarios, they employed sparse coding based on domain-adaptive dictionary learning to enhance the accuracy of the classifier and adaptability of the dictionary. By leveraging the acquired fine-grained attributes of the clothing items and utilizing matching scores, the retrieval results can be re-ranked, optimizing the effectiveness of the item retrieval process.

In order to optimize fashion image retrieval, Goenka et al. [30] proposed the FashionVLP model,which was based on feedback model. This model consists of two parallel blocks: one for processing the reference image and feedback, and the other for processing target images. This approach effectively fuses target image features without relying on text or transformer layers, resulting in increased efficiency in recognition tasks.

The advancements in deep learning have greatly influenced the development of item retrieval by leveraging deep neural network architectures. Huang et al. [31] presented the Dual Attribute-aware Ranking Network (DARN) to capture comprehensive features. During the feature learning phase, semantic attributes and visual similarity constraints are embedded, while addressing inter-domain differences. Ak et al. [32] also utilized the structure of pooling layers and proposed the FashionSearchNet model, which exploits attribute activation maps to learn region-specific attribute representations. This approach enhances the understanding of regions within fashion images and enables the retrieval of fashion items based on specific attributes.

These research showed that item retrieval would no doubt improve the accuracy, efficiency, and effectiveness of fashion detection processing.

III. FASHION SYNTHESIS

A. IMAGE-GUIDED SYNTHESIS

1) OVERVIEW

Due to the advancement of the generative models such as Generative Adversarial Networks (GANs) [1], [33] and Diffusion Models [3], we have witnessed a rapid development in the field of image processing [2], [34]. These sophisticated models have revolutionized the way we generate, manipulate, and enhance images, enabling unprecedented levels of realism and creativity.

Some of these models have been applied in the fashion industry to assist designers in creating and improving fashion items. We categorize them based on the different types of input, including image-guided, sketch-guided, text-guided, and multimodal-guided models. These models provide valuable guidance and inspiration to fashion designers, allowing them to leverage the power of artificial intelligence in their creative process. The following image illustrates the process of fashion synthesis using various inputs.

Most of the image-guided fashion synthesis models are based on fashion style transfer which is a branch of neural style transfer. Fashion style transfer involves taking a fashion style image as the target and generating clothing images that embody the desired fashion styles. The main challenge in Fashion Style Transfer is to preserve the underlying design of the input clothing while seamlessly integrating the desired style patterns. Generative Adversarial Network (GAN) [33] and diffusion models have had significant development and impact in the field of image style transfer.

Goodfellow et al. [33] proposed a new framework for estimating generative models via an adversarial process, in which they simultaneously trained two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. GANs have been successfully applied to image synthesis showing great potential in generating realistic and diverse data. The following works of fashion design is based on the Generative Adversarial Network.

Many of the GAN-based fashion design methods are inspired by the first NST algorithm proposed by Gatys et al. [1], [33]. They construct the content component of the output image by penalizing the difference between high-level representations obtained from the content input images and output images. Additionally, they create the style component by aligning the Gram-based summary statistics of the style input images and output images. As for the task of image-guided fashion synthesis, the models are given appearance input images I_a which represent for texture, pattern, color and so on and structure input images I_s represent for the outline and type of the garments.

$$\begin{aligned} I^* &= \arg\min_{I} \mathcal{L}_{total}(I_a, I_s, I) \\ &= \arg\min_{I} \alpha\mathcal{L}_a(I_a, I) + \beta\mathcal{L}_s(I_s, I), \end{aligned} \tag{1}$$

where L_a evaluates the appearance representation of the appearance input image in relation to the output image, while L_s evaluates the Gram-based structure representation derived from the structure input image in relation to the output image. α and β are the parameters to regulate the balance between the appearance component and structure component in the output.

Denoising diffusion probabilistic model (DDPM) [3] is another generative modeling approach based on diffusion processes, aimed at modeling and generating high-dimensional data. It iteratively updates the conditional density using random noise to progressively generate realistic samples. DDPM introduces denoising priors to further improve the quality of generated samples. It is capable of generating high-quality samples, handling complex data, and finds extensive applications across multiple domains. DDPM has provided new possibilities for the advancement of the fashion design field.

The development of GANs and diffusion models has had a profound impact on the field of image style transfer. They provide powerful tools and methods for designers, artists, and researchers to generate images with unique styles and artistic qualities. These techniques not only offer new possibilities

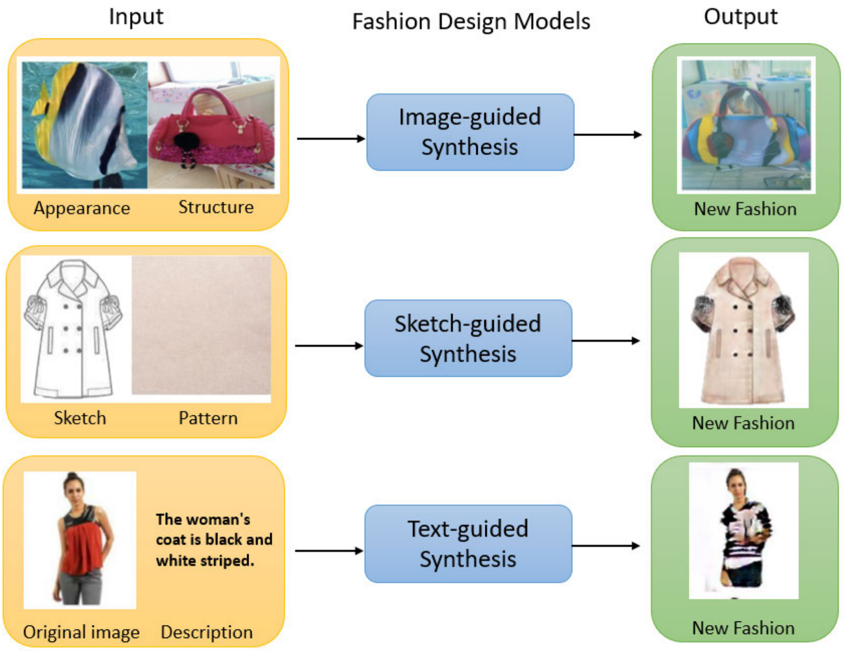


FIGURE 4. Illustration of fashion synthesis with three different types of inputs.

for image generation and style conversion but also hold innovative potential in domains such as digital art, design, and visual effects. Additionally, they have stimulated research and development in computer vision and deep learning, advancing the understanding and exploration of image generation and processing. Therefore, the development of the aforementioned two technologies has created the technological background for the emergence of Image-guided Synthesis Models.

2) DEVELOPMENT

The ground breaking work in this area was proposed by Jiang et al. [11] on the basis of GAN. They are the first one applying artificial intelligence to automatically generate fashion style images. They proposed an end-to-end feed-forward neural network consisting of a fashion style generator G and a discriminator D. The global and patch-based style and content losses, computed by the discriminator, are alternately back-propagated to the generator network for optimization. During the global optimization stage, the shape and design of the clothing are preserved, while the local optimization stage preserves the detailed style patterns.

Although this model has a better effect compared to the previous related global or patch based neural style transfer works, it still has some limitations. When the texture in the input image is not prominent, the output results may be less satisfactory. Also, the network may preserve some of the original colors from the content image and the resolution of the generated clothing is relatively low.

Then, Jiang et al. [35] proposed the FashionG framework also on the basis of GAN for single-style generation and proposed the SC-FashionG framework for for mix-and-match style generation. FashionG includes a generator and a discriminator. It is worth noting that for SC-FashionG, they incorporated a spatial segmentation mask into the input channels to ensure that each style is exclusively assigned to particular regions. This process involves two stages: offline training and online generation. Inputs for the offline stage consist of content images, two style images and two additional channels which are opposite up-down and down-up spatial masks. They are used to guide that the one style is transferred onto the upper part of the output and the other style is transferred onto the bottom part. At this time, the SC-FashionG training framework calculates spatially constrained patch and global reconstruction losses. In the online generation stage, for an input clothing image and an arbitrary spatial mask, they generate outputs with the offline trained generator G. In this way, the framework can generate mix-and-match fashion style output.

Sbai et al. [12] introduced a specific conditioning of GANs on texture and shape elements for generating fashion design images.They tried different Generative Adversarial Networks architectures, novel loss functions to encourage creativity. Moreover, they put together an evaluation protocol associating automatic metrics and human experimental studies to evaluate the results. Although the generated clothing closely resembles designs created by human designers rather than computers, the generation process is relatively random and lacks a specific design direction.

Huang et al. [2] proposed a GAN-based method for multimodal unsupervised image-to-image translation called Multimodal Unsupervised Image-to-image Translation (MUNIT).

It can generate output images with similar content but different styles from the input images in two categories. It was not invented for fashion design, but it can still be applied in this field by training the model with a set of prepared fashion design images. Then, by inputting a design image into the MUNIT model and adjusting the style code, we can achieve transformations to other styles. This allows designers to quickly explore variations in fashion designs with different styles, textures, and patterns.

Apart from this, more projects based on GAN adapted to fashion design area appears. Yan et al. [36] proposed a texture and shape disentangled generative adversarial network (TSD-GAN) to perform well and creative design with the transformation of texture and shape in fashion items.

In the TSD-GAN, an FAEnc module is designed to disentangle the input image features into texture and shape representations. They proposed TMNet and SMNet modules to decompose the texture and shape features into hierarchical representations to capture coarse and fine styles. Their MFGen module aims to utilize these hierarchical representations to synthesize mixed-style fashion items. A Fusion-block module learns the mapping relationship between texture and shape representations. Additionally, a fashion attributes discriminator predicts real-or-fake distributions, while a patch discriminator calculates pixel-wise texture similarity.

Based on shape and texture codes interpolation and principal component analysis [37], the TSD-GAN method assists designers in quickly generating multiple different clothing design options without altering the overall design style and texture. Designers can manipulate the variations in the shape and texture of the clothing by adjusting the weights of principal component vectors, without the need for manual editing or redesigning of the garments. This allows designers to quickly realize their creative ideas and explore a wider range of design possibilities.

Yan et al. [38] proposed a generative adversarial network with heatmap-guided semantic disentanglement (HSD-GAN) to perform an ''intelligent'' design with ''inspiration'' transfer.

Specifically, Texture Brush utilizes two main techniques: heatmap-guided semantic disentanglement and texture brush. The heatmap-guided semantic disentanglement technique decomposes the semantic information in fashion designs or clothing photos into several heatmaps, each representing a specific texture or color. These heatmaps can be used to control texture transfer and color variations. On the other hand, the texture brush technique applies these textures to another clothing photo, enabling texture transfer. They introduced a semantic disentanglement attention-based encoder to capture the most discriminative regions of input items and disentangle the features into attributes and texture. A generator is developed to synthesize mixed-style fashion items by utilizing them. Additionally, they introduced a heatmap-based patch loss to evaluate the visual-semantic matching degree between the input and generated texture.

Yan et al. [39] also proposed a novel intelligent design approach named FadGAN with similar function. FadGAN encodes the source and target images into shared latent vectors and independent attribute vectors using two encoders. During this process, the attribute vectors are separated through an additional attribute encoder, and the inspiration transfer from the source image to the target image is achieved by swapping the attribute vectors. To ensure high-quality fashion attributes in the generated designs, the model utilizes predefined fashion attribute vectors to constrain their consistency with fashion elements. FadGAN consists of an attribute encoder based on Variational Autoencoders (VAEs) [40] and an image generator based on Conditional GANs [41]. During training, it optimizes the entire network by minimizing the reconstruction error and adversarial loss of the image generator and attribute encoder. Ultimately, the model can generate fashion designs with high-quality fashion attributes, aiding designers in faster creation.

To sum up, GAN-based methods can be primely used in fashion design to generate new clothes. However, these methods lack control over the appearance and shape of clothes when transferring from non-fashion domain images.

Before the advent of diffusion models, GANs had already undergone a significant period of development. As a result, the number of models based on diffusion models as the underlying network architecture is relatively fewer. To deal with new fashion design task aimed to transfer a reference appearance image onto a clothing image while preserving the structure of the clothing image, Cao, Chai, et al. [42] presented a reference-based fashion design with structure-aware transfer by diffusion models called DiffFashion. It can semantically generate new clothes from a provided clothing image and a reference appearance image. Specifically, the method separates the foreground clothing using automatically generated semantic masks conditioned on labels. These masks serve as guidance in the denoising process to preserve the structural information. Additionally, a pretrained vision Transformer (ViT) is utilized to guide both the appearance and structure aspects.

Compared to the previous work, DiffFashion can generate more realistic images in the fashion design task. However, due to the randomness of diffusion, the mask cannot guarantee good results every time. Better masks need to be generated to improve the task.

B. SKETCH-GUIDED SYNTHESIS

1) OVERVIEW

Some of the fashion design models are realized by taking sketches as input to create new fashion items or revise existed fashion items. These models can assist fashion designers in quickly and efficiently designing new garments, with excellent human-computer interaction. Designers can fill sketches with different fabric options and make changes to the detailed style of the garments.

2) DEVELOPMENT

Isola et al. [43] introduced a method named conditional adversarial networks for image-to-image translation. It achieves impressive results in various image translation tasks including fashion design by transforming a sketch into a photo. The approach involves training a generator and discriminator network simultaneously. The generator network generates images in the desired domain, while the discriminator network distinguishes between real and generated images. It lays the foundation for the subsequent development of the field of image translation for fashion design.

Xian et al. [44] are the first to examine texture control in the image synthesis area. They proposed an approach called TextureGAN for controlling fashion image synthesis. It allows users to place a texture patch on a sketch at arbitrary locations and scales to control the desired output texture by training the generative network with a local texture loss in addition to adversarial and content loss. The proposed algorithm is able to generate plausible images that are faithful to user controls. But it cannot be used in more complex scenes.

Cui et al. [45] proposed an end-to-end virtual garment display method called FashionGAN based on Conditional GAN. The method requires user input of a fashion sketch and fabric image, enabling quick and automatic display of a virtual garment image that aligns with the input sketch and fabric. Moreover, it can also be extended to contour images and garment images, which further improves the reuse rate of fashion design.

FashionGAN establishes a bijection relationship between fabric and latent vector so that the latent vector can be explained with fabric information. Moreover, some local losses are added to help generate images with stripe texture. The method can indeed achieve impressive results in fashion design, however, the pattern design is relatively limited, and there is still room for improvement.

Dong et al. [46] proposed a novel Fashion Editing Generative Adversarial Network (FE-GAN), which enables users to manipulate fashion images with arbitrary sketches and a few sparse color strokes.

To achieve realistic interactive results, the FE-GAN incorporates a free-form parsing network that predicts the complete human parsing map, guiding fashion image manipulation and crucially contributing to producing convincing results. Additionally, a foreground-based partial convolutional encoder is developed, and an attention normalization layer is designed for the multiple scales layers of the decoder in the fashion editing network. Compared to previous works, the network demonstrates good performance in fashion editing. However, the editing operations themselves have limited functionality, and the quality of the generated results is highly dependent on the input sketch.

Yan et al. [47] introduced a GAN-based AI-driven framework for completing the design of fashion items' sketches.

To incorporate different textures into various designed sketches, conditional feature interaction (CFI) is proposed to learn semantic mapping from the sketch to the texture. Two training schemes are developed, namely end-to-end training and divide-and-conquer training, with the latter demonstrating superior performance in terms of compatibility, diversity, and authenticity. However, the proposed network structure has the following limitations: the generation process is relatively random and lacks controllability, and the model cannot generate images based on color ratios.

C. TEXT-GUIDED SYNTHESIS

1) OVERVIEW

With the development of multimodal machine learning techniques, some AI systems are capable of assisting in fashion design based on textual guidance. Designers simply need to input a textual description of the desired garment into the network architecture to obtain a designed outfit. This method is highly convenient to use, but if more complex images are desired, it places a relatively high demand on the user's descriptive abilities.

2) DEVELOPMENT

Given an input image of a person and a sentence describing a different outfit, GAN based model put forward by Zhu et al. [48] can dress the person as desired while preserving their pose. They decomposed the complex generative process into two conditional stages. In the first stage, they generated a plausible semantic segmentation map that aligns with the wearer's pose as a latent spatial arrangement. They incorporated an effective spatial constraint to guide the generation of this map. In the second stage, they employed a generative model with a newly proposed compositional mapping layer to render the final image, considering precise regions and textures conditioned on the generated semantic segmentation map.

However, the results generated are limited by the current database they adopted, for the training set contains images mostly with a plain background.

Günel et al. [49] proposed a novel approach called FiLMedGAN, which utilizes feature-wise linear modulation (FiLM) to establish a connection between visual features and natural language representations, enabling their transformation without relying on additional spatial information. The approach adopts a GAN-based architecture that enables users to edit outfit images by inputting different descriptions to generate new outfits. The FiLMedGAN model, incorporating skipping connections and FiLMed residual blocks, achieves excellent performance and meets the desired objectives.

Zhou et al. [50] presented a method to manipulate the visual appearance like pose and attribute of a person's image according to natural language descriptions. They presented a novel two-stage pipeline to achieve it. The pipeline first learns to infer a reasonable target human pose based on the description, and then synthesizes an appearance transferred person image according to the text in conjunction with the target pose. However, the project mainly showcases the functionality of changing the color and design of clothing, which is relatively simple and limited in scope.

There is a task called composing text and image to image retrieval (CTI-IR), which aims to retrieve images relevant to a query image based on text descriptions of desired modifications. To deal with it, Zhang et al. [51] proposed an end-to-end trainable network based on GAN for simultaneous image generation and CTI-IR. The model presented in this study learns generative and discriminative features for the query image through joint training of a generative model and a retrieval model. It automatically manipulates the visual features of the reference image based on the text description using adversarial learning between the synthesized and target images. Global-local collaborative discriminators and attention-based generators are leveraged to focus on both global and local differences between the query and target images.

As a result, the model enhances semantic consistency and fine-grained details in the generated images, which can be utilized for interpretation and empowerment of the retrieval model. The network combines the fields of retrieval and image generation to achieve the effect of fashion design.

D. MULTIMODAL-GUIDED SYNTHESIS

With the advancements in multimodal technologies in the field of AI, fashion design tasks have witnessed significant enhancements, enabling designers to leverage multimodal inputs for enhanced creativity and convenience. The three main modalities of control signals for conditional image synthesis: textual controls, visual controls including image and sketch input, and preservation controls refers to complete the missing parts in an image.

Zhang et al. [13] proposed a novel two-stage architecture called M6-UFC, which aims to unify multiple multimodal controls in a universal form for conditional image synthesis. Non-auto regressive generation (NAR) is utilized to improve inference speed, enhance holistic consistency, and support preservation controls. Additionally, a progressive generation algorithm based on relevance and fidelity estimators is designed by the authors to ensure relevance and fidelity.

IV. FASHION RECOMMENDATION

After fashion detection and synthesis, fashion recommendation has emerged as a developing and challenging area. Customers now require the ability to discover desired products tailored to various situations. Over the past fifteen years, an increasing number of highly efficient computer vision algorithms have been developed to address this demand in the market.

Based on user requirements, recommender systems [52], [53] can be categorized into two types: task-based and input-based recommender systems [4]. Task-based recommender systems are designed for specific scenarios that users intend to engage in, such as parties or trips. These systems assist users in selecting fashion items or generating matching designs that align with the desired scenario. On the other hand, input-based recommender systems cater to users who already possess certain items and seek recommendations for matching items from their existing collection. These systems aid users in selecting a complementary item that pairs well with their current possessions.

A. TASK-BASED RECOMMENDER SYSTEMS

In the realm of fashion recommendations, users frequently make choices based on their individual needs, such as specific activities they plan to participate in.

Shen et al. [54] have pioneered the development of a Scenario-Oriented recommendation system. This system utilizes the open mind common sense (OMCS) [55], a comprehensive knowledge corpus encompassing people's everyday common sense. It represents the first technology that offers product recommendations based on real-world user scenarios, encompassing a wide array of themes. Leveraging contextual information and potential product associations, this system assists users in easily identifying the most suitable product, even when they are unsure about specific details. Jagadeesh et al. [56] have introduced two categories of recommenders: deterministic recommenders (DFR) and stochastic recommenders (SFR). Their approach involves extracting fashion insights from vast amounts of online fashion images and their accompanying rich metadata. These recommenders enable the identification of valuable fashion-related patterns and trends.

The previously mentioned recommender methods were considered relatively basic, lacking the utilization of neural networks. As consumer needs continue to grow, more advanced and effective recommender models have been proposed.

Huynh et al. [57] introduced a groundbreaking unsupervised learning approach called complementary recommendation using adversarial feature transform (CRAFT). CRAFT builds upon the principles of GANs [33]. However, unlike direct image synthesis, CRAFT focuses on training in the feature space. The genetic converter within CRAFT is capable of generating diverse characteristics through the utilization of random input vectors. The transformed feature vector is then employed to recommend images based on their nearest neighbors in the feature space. By learning the joint distribution of co-occurring visual objects in an unsupervised manner, CRAFT is applied to visual complementary recommendation. Remarkably, this approach does not necessitate the presence of annotations or labels to indicate complementary relationships.

Previous recommender systems often relied on recommending fashion items similar to a user's previous purchases, resulting in a lack of item diversity. To address this challenge, De Divitiis et al. [58] proposed a self-supervised contrastive learning model known as memory augmented neural networks (MANNs). This approach tackles the issue by combining color and shape feature disentanglement. It incorporates external memory modules that store pairing modalities between different types of clothing, such as tops and bottoms. By addressing issues arising from imbalanced data distribution, compact and representative memories are obtained. These memories are then used to expand the common controller loss, facilitating the training of the memory modules. The usage of MANNs with disentangled features and memory augmentation enables the recommender system to recommend a more diverse range of fashion items to users. This approach enhances the overall recommendation quality and overcomes the limitations of previous systems.

Ye et al. [59] introduced a scene-aware fashion recommender system (SAFRS) that takes into account the context of recommendations, which accurately selects outfits based on the scene context provided and completes diverse scene-aware fashion recommendation tasks.

B. INPUT-BASED RECOMMENDER SYSTEMS

Users not only choose fashion items with the use of scenarios, they will also choose according to their own conditions. If they choose in the fashion items they have, the input-based algorithms will come in handy at this time. For early work, Iwata et al. [60] proposed a probabilistic topic model to recommend tops for bottoms by learning information about coordinates from visual features in each fashion item. Given a photo of a fashion item (tops or bottoms) used as a query, the recommender system first detects the face region, and determines the top and bottom regions by assuming that they are were divided in a certain proportion. The segmentation method is rough and far from reality. To meet the user requirements, new input-based algorithms have emerged in recent years. More and more innovative network architectures or data processing methods have been used to better adapt to the needs of users.

Lu et al. [61] proposed a learningable personalized anchor embedding (LPAE) method for personalized clothing recommendation and new user analysis. This model uses stacked self-attention mechanism to encode clothing into compact embedded to capture the high-level relationship between fashion items. It uses a set of anchors to model the fashion preferences of each user, and calculate the similarity between the preference score based on the embedded clothing and the user anchor, where the similar encoded fashion items can be retrieved according to the distance in space. This method is in an advanced position in ordinary settings and cold start settings.

Delmas et al. [62] proposed attention-based retrieval with text-explicit matching and implicit similarity (ARTEMIS), a new method for image search with free-form text modifiers. It has two modules focusing on how potential targets fit the textual modifier, and comparing potential target images to the reference image assisted by the text, respectively.

Among numerous structures, some methods have innovated around some core elements. Below are some representative network architectures or methods.

1) METHODS BASED ON KNOWLEDGE GRAPH

Zhan et al. [63] proposed an end-to-end personalized outfit recommender system (A^3-FKG), which investigates the usage of knowledge graph in capturing the connectivity between entities and exploits the complementary benefits of the multimodal information. It includes two-level attention modules, corresponding to user-specific relation-aware and target-aware networks, which fits knowledge graph into the recommender system in the fashion domain.

Dong et al. [64] proposed a new designer-oriented recommender system using knowledge graph to support personalised fashion design, which provides a feedback and self-adjustment mechanism that can the users' perceptual feedback and adjust its knowledge base automatically.

2) TEXTUAL FEATURES

Facing the challenges of visual understanding and visual matching, Lin et al. [65] proposed a co-supervision learning framework, namely FARM. It captures aesthetic characteristics with the supervision of generation learning, and includes a layer-to-layer matching mechanism to evaluate the matching score. FRAM can deal with the situation that given an image of top and a query bottom item in vector description, where it can generate a image of matching bottom.

V. DISCUSSION

A. MARKETING AND CONVENIENCE

AI-generated content (AIGC) [66], [67] is revolutionizing industries, fashion design included. The market potential is enormous, especially in fields like fashion. AIGC meets the demand for captivating content, such as product descriptions and trend analyses, sparking interest among designers and businesses.

AIGC's cost-cutting prowess is impressive. Traditional content creation requires various specialists, but AIGC automates processes, leading to substantial savings. This efficiency lets brands allocate resources strategically. Beyond cost, AIGC brings unprecedented convenience. It generates content on-demand, eliminating delays seen with human content creation. This ensures consistent, timely material flow-a boon for fashion brands chasing real-time trends.

In fashion design, AI complements human designers. It offers insights from data, predicts trends, and automates design elements. This blend enhances designers' capabilities, combining creativity with data-driven insights. In short, AIGC's success is transformative. It reshapes creativity, efficiency, and convenience across markets, particularly in fashion. The partnership of human and AI promises an innovative future.

B. CHALLENGES AND FUTURE TRENDS

With the development in recent years, we are still facing some challenges in fashion design, especially in terms of model training and user requirements. Here are a few prominent challenges listed.

- **Data Quality and Diversity:** Ensuring access to high-quality and diverse fashion datasets remains a challenge. Developing more comprehensive and representative data sets covering different styles and cultures will improve the performance and inclusiveness of AI models in Fashion design and recommendations. Currently, many datasets are not open-source, leading to situations where resources cannot be shared. This poses challenges for design.

- **Real-Time Fashion Synthesis:** Improving the efficiency and speed of fashion synthesis models is crucial for real-time design applications. Optimizing algorithms and utilizing hardware advancements can enable designers to interactively explore and iterate design options, thereby simplifying the design process.
- **Multimodal Integration:** Further research is needed to strengthen the integration of multimodal inputs in fashion design and recommender systems. Creating seamless interaction between text, images, sketches, and other forms will enable designers to express their ideas more effectively, resulting in more accurate and personalized output.
- **Interpretability and Explainability:** As artificial intelligence models become increasingly complex, understanding their decision-making process and providing explanations for their recommendations becomes crucial. Developing interpretable and explainable artificial intelligence systems in fashion design and recommendation will increase trust and enable designers and users to understand and modify the generated output.
- **Combination with Industry:** AIGC has brought a transformative shift to designers' roles, demanding adaptation. They must embrace data-driven insights and collaborate with AI to stay trend-aligned. Combining human creativity with AI-generated content sparks unique designs and enables rapid prototyping for trend responsiveness.

The future of fashion design is heading towards a convergence of enhanced personalization and seamless multimodal interfaces. With development in fashion design, personalized experiences will take center stage, as AI systems leverage user preferences, body measurements, and style choices to provide tailored recommendations. Simultaneously, the integration of various input modalities such as text, images, sketches, and virtual reality will enable designers and users to communicate their ideas effortlessly. This fusion of enhanced personalization and seamless multimodal interfaces will empower fashion designers and consumers to co-create unique and personalized designs that perfectly align with individual tastes and preferences. It will revolutionize the fashion industry, fostering a deeper level of engagement, satisfaction, and creativity in the design process. Furthermore, augmented reality is gradually applied in fashion industry to enhance the user experiences [68], where Users can perform some virtual try-on, rather than just viewing items.

VI. CONCLUSION

The survey covered different topics related to fashion design, including fashion detection, synthesis, recommendation, and the use of multimodal inputs.

In the area of fashion detection, computer vision algorithms have been developed to accurately identify fashion items in images or videos. These algorithms help in analyzing fashion trends, understanding consumer preferences, and providing valuable insights to fashion designers and retailers.

Fashion synthesis techniques, such as GAN-based models, have emerged as powerful tools for generating new clothing designs. These models can disentangle and manipulate shape, texture, and style representations, allowing designers to explore a wide range of design possibilities quickly and efficiently. Diffusion models, with denoising priors, have also shown promise in generating high-quality and realistic samples.

Fashion recommendation systems play a crucial role in assisting customers in finding the products they are looking for based on different situations and conditions. Traditional recommendation methods have been enhanced with the integration of computer vision algorithms, probabilistic topic models, and deep learning architectures. These systems can provide personalized recommendations, considering factors such as user preferences, contextual information, and complementary relationships between fashion items.

The survey also highlighted the importance of multimodal inputs in fashion design. Models that utilize textual, visual, and preservation controls have been developed to enable designers to create new garments or revise existing ones more efficiently. Sketch-guided models allow designers to input sketches and generate fashion items accordingly, while text-guided models assist in designing outfits based on textual descriptions.

In conclusion, the survey demonstrates the significant progress made in the application of AI and machine learning techniques in fashion design. These advancements have resulted in more efficient and creative design processes, personalized recommendations, and improved user experiences. As the field continues to evolve, further research and development in areas such as self-attention mechanisms, knowledge graphs, and textual features are expected to enhance the capabilities of fashion design and recommendation systems even further.

ACKNOWLEDGMENT

(Ziyue Guo, Zongyang Zhu, and Yizhi Li contributed equally to this work.)

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[2] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[4] Y. Deldjoo, F. Nazary, A. Ramisa, J. Mcauley, G. Pellegrini, A. Bellogin, and T. Di Noia, "A review of modern fashion recommender systems," 2022, *arXiv:2202.02757*.

[5] L. Regenwetter, A. H. Nobari, and F. Ahmed, "Deep generative models in engineering design: A review," *J. Mech. Des.*, vol. 144, no. 7, Jul. 2022, Art. no. 071704.

[6] M.-F. de-Lima-Santos and W. Ceron, "Artificial intelligence in news media: Current perceptions and future outlook," *Journalism Media*, vol. 3, no. 1, pp. 13–26, Dec. 2021.

[7] K. Hara, V. Jagadeesh, and R. Piramuthu, "Fashion apparel detection: The role of deep convolutional neural network and pose-dependent priors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[8] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 229–245.

[9] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3570–3577.

[10] Z. Li, Y. Li, W. Tian, Y. Pang, and Y. Liu, "Cross-scenario clothing retrieval and fine-grained style recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2912–2917.

[11] S. Jiang and Y. Fu, "Fashion style generator," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3721–3727.

[12] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie, "Design: Design inspiration from generative networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018.

[13] Z. Zhang, J. Ma, C. Zhou, R. Men, Z. Li, M. Ding, J. Tang, J. Zhou, and H. Yang, "M6-UFC: Unifying multi-modal controls for conditional image synthesis via non-autoregressive generative transformers," 2021, *arXiv:2105.14211*.

[14] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 4, pp. 1–41, 2021.

[15] X. Gu, F. Gao, M. Tan, and P. Peng, "Fashion analysis and understanding with artificial intelligence," *Inf. Process. Manag.*, vol. 57, no. 5, Sep. 2020, Art. no. 102276.

[16] S. O. Mohammadi and A. Kalhor, "Smart fashion: A review of AI applications in the fashion & apparel industry," 2021, *arXiv:2111.00905*.

[17] C.-H. Lee and C.-W. Lin, "A two-phase fashion apparel detection method based on YOLOv4," *Appl. Sci.*, vol. 11, no. 9, p. 3782, Apr. 2021.

[18] V. Rosenfeld, "Two-stage template matching," *IEEE Trans. Comput.*, vol. C-26, no. 4, pp. 384–393, Apr. 1977.

[19] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.

[20] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4271–4280.

[21] C.-Q. Huang, J.-K. Chen, Y. Pan, H.-J. Lai, J. Yin, and Q.-H. Huang, "Clothing landmark detection using deep networks with prior of key point associations," *IEEE Trans. Cybern.*, vol. 49, no. 10, pp. 3744–3754, Oct. 2019.

[22] S. Lee, S. Oh, C. Jung, and C. Kim, "A global-local embedding module for fashion landmark detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3153–3156.

[23] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1028–1040, May 2015.

[24] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016.

[25] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.

[26] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 115–127, Jan. 2017.

[27] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 4814–4821.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[29] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-CNN meets KNN: Quasi-parametric human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1419–1427.

[30] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan, "FashionVLP: Vision language transformer for fashion retrieval with feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14085–14095.

[31] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1062–1070.

[32] K. E. Ak, A. A. Kassim, J. H. Lim, and J. Y. Tham, "Learning attribute representations with localization for flexible fashion search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7708–7717.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[35] S. Jiang, J. Li, and Y. Fu, "Deep learning for fashion style generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4538–4550, Sep. 2022.

[36] H. Yan, H. Zhang, J. Shi, J. Ma, and X. Xu, "Toward intelligent fashion design: A texture and shape disentangled generative adversarial network," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 3, pp. 1–23, Aug. 2023.

[37] K. Pearson, "Principal components analysis," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 6, no. 2, p. 559, Jan. 1901.

[38] H. Yan, H. Zhang, J. Shi, and J. Ma, "Texture brush for fashion inspiration transfer: A generative adversarial network with heatmap-guided semantic disentanglement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2381–2395, May 2023.

[39] H. Yan, H. Zhang, J. Shi, J. Ma, and X. Xu, "Inspiration transfer for intelligent design: A generative adversarial network with fashion attributes disentanglement," *IEEE Trans. Consum. Electron.*, early access, Mar. 10, 2023, doi: 10.1109/TCE.2023.3255831.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[42] S. Cao, W. Chai, S. Hao, Y. Zhang, H. Chen, and G. Wang, "DiffFashion: Reference-based fashion design with structure-aware transfer by diffusion models," 2023, *arXiv:2302.06826*.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[44] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8456–8465.

[45] Y. R. Cui, Q. Liu, C. Y. Gao, and Z. Su, "FashionGAN: Display your fashion design using conditional generative adversarial nets," *Comput. Graph. Forum*, vol. 37, no. 7, pp. 109–119, 2018.

[46] H. Dong, X. Liang, Y. Zhang, X. Zhang, X. Shen, Z. Xie, B. Wu, and J. Yin, "Fashion editing with adversarial parsing learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8117–8125.

[47] H. Yan, H. Zhang, L. Liu, D. Zhou, X. Xu, Z. Zhang, and S. Yan, "Toward intelligent design: An AI-based fashion designer using generative adversarial networks aided by sketch and rendering generators," *IEEE Trans. Multimedia*, vol. 25, pp. 2323–2338, 2022.

[48] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own Prada: Fashion synthesis with structural coherence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1689–1697.

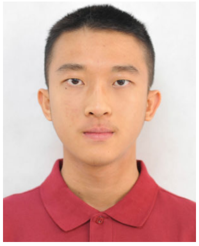
[49] M. Günel, E. Erdem, and A. Erdem, "Language guided fashion image manipulation with feature-wise transformations," 2018, *arXiv:1808.04000*.

[50] X. Zhou, S. Huang, B. Li, Y. Li, J. Li, and Z. Zhang, "Text guided person image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3658–3667.

[51] F. Zhang, M. Xu, and C. Xu, "Tell, imagine, and search: End-to-end learning for composing text and image to image retrieval," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 18, no. 2, pp. 1–23, 2022.
[52] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 197–206.
[53] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, Nov. 2019, pp. 1441–1450.
[54] E. Shen, H. Lieberman, and F. Lam, "What am I Gonna wear? Scenario-oriented recommendation," in *Proc. 12th Int. Conf. Intell. User Interfaces*, Jan. 2007, pp. 365–368.
[55] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *On the Move to Meaningful Internet Systems 2002, CoopIS, DOA, and ODBASE*. Berlin, Germany: Springer, 2002, pp. 1223–1237.
[56] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large scale visual recommendations from street fashion images," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1925–1934.
[57] C. P. Huynh, A. Ciptadi, A. Tyagi, and A. Agrawal, "CRAFT: Complementary recommendation by adversarial feature transform," Tech. Rep., 2018.
[58] L. De Divitiis, F. Becattini, C. Baecchi, and A. Del Bimbo, "Disentangling features for fashion recommendation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1s, pp. 1–21, Feb. 2023.
[59] T. Ye, L. Hu, Q. Zhang, Z. Y. Lai, U. Naseem, and D. D. Liu, "Show me the best outfit for a certain scene: A scene-aware fashion recommender system," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 1172–1180.
[60] T. Iwata, S. Watanabe, and H. Sawada, "Fashion coordinates recommender system using photographs from fashion magazines," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–6.
[61] Z. Lu, Y. Hu, Y. Chen, and B. Zeng, "Personalized outfit recommendation with learnable anchors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12717–12726.
[62] G. Delmas, R. Sampaio de Rezende, G. Csurka, and D. Larlus, "ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity," 2022, *arXiv:2203.08101*.
[63] H. Zhan, J. Lin, K. E. Ak, B. Shi, L.-Y. Duan, and A. C. Kot, "A^3-FKG: Attentive attribute-aware fashion knowledge graph for outfit preference prediction," *IEEE Trans. Multimedia*, vol. 24, pp. 819–831, 2021.
[64] M. Dong, X. Zeng, L. Koehl, and J. Zhang, "An interactive knowledge-based recommender system for fashion product design in the big data environment," *Inf. Sci.*, vol. 540, pp. 469–488, Nov. 2020.
[65] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. De Rijke, "Improving outfit recommendation with co-supervision of fashion generation," in *Proc. World Wide Web Conf.*, 2019, pp. 1095–1105.
[66] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.
[67] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin, "AI-generated content (AIGC): A survey," 2023, *arXiv:2304.06632*.
[68] C. Jayamini, A. Sandamini, T. Pannala, P. Kumarasinghe, D. Perera, and K. Karunanayaka, "The use of augmented reality to deliver enhanced user experiences in fashion industry," in *Proc. IFIP Conf. Human-Comput. Interact.* (Lecture Notes in Computer Science), vol. 12936, 2021, pp. 1–11.

ZIYUE GUO was born in Hebei, China, in 2002. He is currently pursuing the B.E. degree majoring in electrical engineering with Zhejiang University, Zhejiang, China, and the B.S. degree with the University of Illinois Urbana-Champaign, IL, USA. He embarked on his academic journey with the University of Illinois Urbana-Champaign as a half-year exchange student, in August 2022.

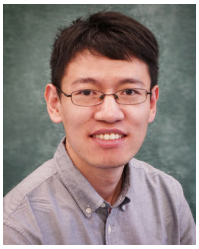
ZONGYANG ZHU was born in Hebei, China, in 2001. He is currently pursuing the B.E. degree majoring in electrical engineering with Zhejiang University, Zhejiang, China, and the B.S. degree with the University of Illinois Urbana-Champaign, IL, USA. He came to the University of Illinois Urbana-Champaign as an exchange student, from August 2022 to August 2023.



YIZHI LI was born in Guangdong, China. He is currently pursuing the B.E. degree majoring in computer engineering with Zhejiang University, China, and the B.S. degree with the University of Illinois Urbana-Champaign, USA. He embarked on his academic journey with the University of Illinois Urbana-Champaign as a half-year exchange student, in August 2022.

SHIDONG CAO received the B.E. degree from the Beijing University of Posts and Telecommunications, China. He is currently pursuing the M.S. degree with Zhejiang University—University of Illinois Urbana-Champaign Institute (ZJUI), Zhejiang University. His research interests include machine learning and computer vision.

HANGYUE CHEN received the master's degree in industrial design engineering from Zhejiang University, in 2010. He is currently a Lecturer with the Department of Digital Media Art, Hangzhou Dianzi University. His current research interests include product digital design and human–computer interaction design.



GAOANG WANG (Member, IEEE) received the B.S. degree from Fudan University, in 2013, the M.S. degree from the University of Wisconsin-Madison, in 2015, and the Ph.D. degree from the Information Processing Laboratory, Electrical and Computer Engineering Department, University of Washington, in 2019. After that, he joined Megvii U.S. Office, in July 2019, as a Research Scientist working on multi-frame fusion. He joined the international campus of Zhejiang University as an Assistant Professor, in September 2020. He is currently an Adjunct Assistant Professor with the University of Illinois Urbana-Champaign Institute. Then, he joined Wyze Laboratories, in November 2019, working on deep neural network design for edge-cloud collaboration. His research interests include computer vision, machine learning, artificial intelligence, including multi-object tracking, representation learning, and active learning. He has published papers in many renowned journals and conferences, including IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Vehicular Technology, CVPR, ICCV, ECCV, ACM MM, and IJCAI.

• • •