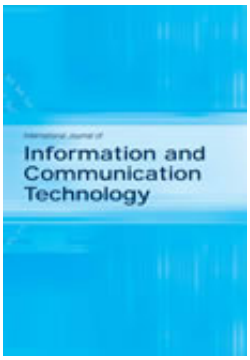
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Adaptive perception enhancement-based virtual try-on technology for accessories with the assistance of artificial intelligence

Ya Liu, Hui Xiao

DOI: 10.1504/IJICT.2025.10071017

Article History:

Received:	27 March 2025
Last revised:	09 April 2025
Accepted:	10 April 2025
Published online:	27 May 2025

Adaptive perception enhancement-based virtual try-on technology for accessories with the assistance of artificial intelligence

Ya Liu* and Hui Xiao

College of Visual Arts,
Hunan Mass Media Vocational and Technical College,
Changsha 410100, China
Email: liuya1119@163.com
Email: fenglingxxx@163.com
*Corresponding author

Abstract: The image quality generated by the traditional virtual try-on technique for non-heritage accessories is poor, to address this problem, this paper firstly designs a convolutional neural network that adaptively adjusts the feature extraction strategy, and adopts an improved generative adversarial network to generate a primary feature map. Then the background noise of the primary feature map is suppressed based on multi-scale attention, and an adaptive perceptual enhancement module is designed to weight the features at different locations in the feature map to strengthen the representation of important features. Finally, the primary feature maps are perspective corrected and downscaled using multi-scale weights to enable the network to generate high-quality images of non-heritage accessory try-on. Experimental results on UNESCO and VITON datasets show that the structural similarity (SSIM) of the suggested method improves 3.45–26.76% compared to benchmark methods, which can effectively improve the quality of image generation.

Keywords: virtual try-on; convolutional neural network; generative adversarial network; GAN; multiscale attention; adaptive perceptual enhancement.

Reference to this paper should be made as follows: Liu, Y. and Xiao, H. (2025) 'Adaptive perception enhancement-based virtual try-on technology for accessories with the assistance of artificial intelligence', *Int. J. Information and Communication Technology*, Vol. 26, No. 14, pp.104–119.

Biographical notes: Ya Liu received her Master's degree from Gwangju Women's University, South Korea, in 2014. She is currently a Lecturer at the Hunan Mass Media Vocational and Technical College. Her research interests include character image design and virtual design.

Hui Xiao received her Master's degree from Hunan Normal University in 2010. She is currently a Professor at the Hunan Mass Media Vocational and Technical College. Her research interests encompass jewellery design and AI-assisted interaction.

1 Introduction

Intangible cultural heritage, as a treasure of human civilisation, carries rich historical and cultural information and national wisdom. As an important carrier of non-heritage culture, non-heritage accessories attract more and more people's attention with their unique shape, exquisite craftsmanship and profound cultural connotation (Cai et al., 2024). However, due to the complexity and high cost of the production process of non-heritage accessories, as well as geographical restrictions and other factors, it is difficult for people to experience the effect of wearing them in person (Xu et al., 2021). In recent years, virtual try-on technology, as an emerging human-computer interaction technology, has provided new ideas for the digital display and experience of non-heritage accessories. However, traditional virtual try-on technology often suffers from a lack of realism and poor interaction experience, which makes it difficult to meet the user's demand for authenticity and immersion in non-heritage accessories try-on (Zhang et al., 2019). As the artificial intelligence explosively developing, 2D image-based virtual fitting methods have become mainstream. In contrast to the complex 3D modelling, the 2D approach transforms virtual fitting into an image generation problem, and how to efficiently achieve the virtual fitting generation of accessories with the assistance of artificial intelligence is a challenging topic (Islam et al., 2024).

As an important bridge connecting traditional culture and modern technology, the virtual fitting technology of non-legacy accessories has received extensive attention from scholars, and the current research in this field mainly focuses on the virtual fitting technology (Hauswiesner et al., 2013). Jiang et al. (2023) used the constraint relationship between curvature for feature point matching between the human body model and the accessory model to achieve the fitting of the human body and clothing, but the fit was not high. Dayik et al. (2016) proposed a virtual fitting method based on Kinect (Xie and Liao, 2014), which solves the skeleton transformation matrix of Kinect to simulate the collision between clothing fabric and human body in real time, and combines somatosensory interaction and real-time rendering to enhance the virtual fitting experience. Sabina et al. (2014) provide methods to build 3D dress fitting with an interactive virtual store and create virtual models using standard sizing parameters, but in poor real-time. Three-dimensional virtual fitting can realise accurate physical simulation. However, it often requires expensive 3D scanning equipment, and a large amount of modelling work not adapted to the current real-time interactive application scenarios.

With the explosive growth of deep learning technology, 2D virtual try-on technology has become a research hotspot. Image-based virtual fitting is addressed as an image generation problem, which greatly reduces human and material resources and improves the real-time nature of virtual fitting. Zhou et al. (2024) adjusted the diffusion model according to image information such as Canny edge map, depth map, and human body key point map to control the human body posture, edge features, and front/back position relationship of the generated image, which provided a new idea for realising fast virtual fitting. Wang et al. (2022) proposed a virtual fitting method based on generative adversarial networks (GAN), which synthesises the flat clothes twisted and deformed into the corresponding region of the model image to realise virtual fitting, but the quality of image generation is poor. Ishikawa and Ikenaga (2022) used a two-channel CNN to express the regularised masks for each accessory category as linear combinations of learnable mask templates to predict more accurate masks. Chou et al. (2024) proposed

contextual CNN to solve the human try-on parsing problem, and the framework captures cross-layer, global image-level and local hyperpixel semantic information well enough to enable pixel-level classification.

A large amount of background noise is often generated in virtual try-on images to interfere with the generation of high-quality images, so perceptual enhancement methods are needed to strengthen foreground information and weaken useless features to improve performance. The method mainly designs the spatial attention (SAM) module (Zhang et al., 2021), which filters the image background noise through the attention map of the accessory virtual try-on. Ye et al. (2024) combined diffusion model and attention mechanism to implement virtual fitting technique, and the attention mask generated by the attention scale branch to segment the accessory region and background, but the computational effort is large. Hu et al. (2022) used the proposed dual path SAM to generate the binary and preliminary density maps of segmented foregrounds for multiplication to reduce the effect of background noise on the results and synthesised the final fitting image using the U-Net generator (Siddique et al., 2021) to improve the generation results.

Through the research and analysis of virtual wearing technology, it is known that the existing virtual trying on technology leads to insufficient user experience due to the poor quality of image generation. To address the above issues, in this paper, the virtual try-on technology of accessories based on adaptive perception enhancement with the assistance of artificial intelligence. The main innovations of this technology are summarised in the following four aspects.

1 Adaptive null CNN is designed to adaptively adjust the feature extraction strategy according to different non-heritage accessories trying on scenarios, and a multi-task discriminator is designed on the basis of conditional GAN, which enables the network to capture different levels of the primary feature maps through the multi-scale strategy, so as to generate more realistic feature maps.

2 The background noise rejection (BNR) module is designed based on efficient channel attention (ECA) and multi-scale fusion. The effective information of different scales of the primary feature map is extracted by BNR and the effective feature information is fully expressed for better noise suppression.

3 Adaptive perceptual enhancement module is designed to weight the features at different locations in the feature map according to the weight information in the feature map to enhance the expression of important features. Multi-scale weights are utilised to perspective correct and downscale the primary feature maps, so that the network generates high-quality images of non-heritage accessories trying on.

4 The experimental results on the public datasets UNESCO and VITON show that the SSIM of the proposed method is 0.809 and 0.916 respectively, which significantly improves the quality of the generation of the images of non-heritage accessories trying on, and provides an innovative path for the digital preservation and inheritance of non-heritage culture.

2 Relevant technologies

2.1 Generative adversarial network

GANs generate data that is virtually indistinguishable from real data through an adversarial process, and consist of a generator and a discriminator, which compete with each other during training to continuously improve performance (Creswell et al., 2018). The GAN training process is a two-player zero-sum game of min-max. The generator tries to maximise the probability that the discriminator will misclassify the generated data, while the discriminator tries to minimise this probability and accurately distinguish between true and false data, as shown below:

$$\min \max V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

where x is the real data, z is the input noise to the generator, $D(x)$ is the probability that the discriminator evaluates the data as real, and $G(z)$ is the fake data generated by the generator based on the noise.

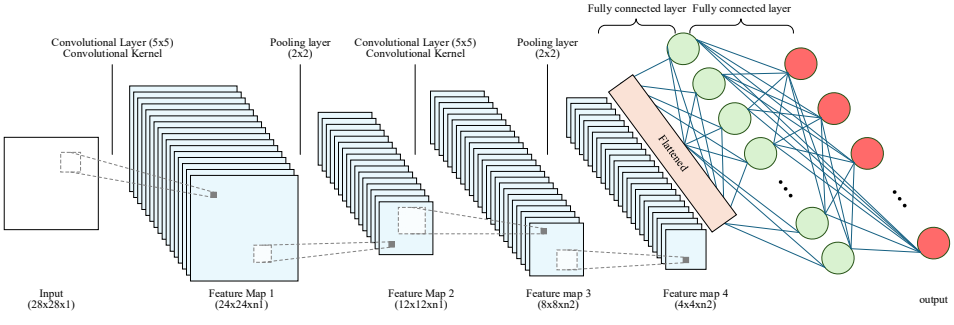
Conditional GAN (CGAN) is an extended form of GAN that controls specific attributes of the generated data by introducing additional conditional information (Douzas and Bacao, 2018) in the inputs of the generator and discriminator, so that the generated data is not only of high quality but also able to fulfil specific conditions or requirements. The objective function of CGAN is as follows, where y is the conditional information.

$$\min \max V(G, D) = E_{x \sim p_{data}(x)}[\log D(x|y)] + E_{x \sim p_z(z)}[\log(1 - D(G(z|y)))] \quad (2)$$

2.2 Convolutional neural network

The processing of images can be viewed as a large matrix operation, and the convolutional and pooling layers of CNNs are more suited to the operation on images than the full connectivity of feed-forward neural networks (Tripathi, 2021). CNN has become the most widely used deep learning network in the image field, which has greatly promoted the growth of computer vision. Non-legacy accessories contain rich and unique textures, patterns and other detailed features, CNN through the convolutional kernel in the convolutional layer, can be able to locally perceive the input non-legacy accessories image data, and automatically extract these complex and subtle features.

Figure 1 Convolutional neural network architecture (see online version for colours)



As shown in Figure 1, CNN is mainly composed of input layer, convolutional layer, downsampling level, fully connected level, output level, and activation function. Assuming that the height and width of the feature map are H and W, respectively, the feature map output by CNN is as follows:

$$Output_H = \left\lfloor \frac{H + 2P - F_H}{s} + 1 \right\rfloor \tag{3}$$

$$Output_w = \left\lfloor \frac{W + 2P - F_W}{s} + 1 \right\rfloor \tag{4}$$

where F_H is the height of the convolution kernel, F_W is the width of the convolution kernel, P is the padding, S is the step size, and $\lfloor\ \rfloor$ is the downward rounding. The fully-connected layer uses matrix multiplication to transform the feature dimensions into one-dimensional feature vectors of fixed length, as shown below:

$$f(x) = W \times x + b \tag{5}$$

where $f(x)$ is the output, W is the weight, b is the bias, and x is the input feature.

2.3 Spatial attention mechanism

The SAM explicitly models the spatial relationship between pixels or regions (such as the relevance of adjacent pixels), which is suitable for structured data such as images and videos. In contrast, the ordinary attention mechanism indirectly introduces spatial information through position encoding. SAM dynamically adjusts the weights of different spatial locations in the feature map by learning the features of the input data, thus realising the enhancement of critical regions and the suppression of irrelevant regions, which is essentially a process of adaptive perceptual enhancement (Wang et al., 2024). Assuming that the input SAM feature map dimension is $C \times H \times W$, based on the channel dimension, we sequentially compute the maximum value $F^s_{\max}$ and the average value pooling F^s_{avg}, and concat the two channel dimension features to obtain the feature map of $2 \times H \times W$. Then it is downscaled to a feature map of $1 \times H \times W$ by a convolutional layer, and then the attention weights are obtained by a sigmoid activation function. Finally, this weight is dot-multiplied with the input feature map to obtain the feature map containing spatial location information, the equation is as follows:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \end{aligned} \tag{6}$$

where $M_s(F)$ is the output of SAM, F is the input feature map, $f^{7\times7}$ is the convolution operation with 7×7 convolution kernel, and σ is the sigmoid operation.

3 Primary intangible cultural heritage accessory feature map generation based on generative adversarial network

3.1 Adaptive null convolution-based feature extraction for non-legacy accessories

To address the limitation of virtual try-on in dealing with complex texture representation and feature interaction of non-heritage accessories, a multi-task discriminator-based method for generating primary feature maps of non-heritage accessories try-on is proposed, which lays the foundation for generating high-quality virtual try-on results of non-heritage accessories. Adaptive cavity convolution (ACNN) module is firstly designed to adaptively adjust the feature extraction strategy according to the different input images, so as to better extract the effective features in different NRL accessory try-on scenarios. Then a multi-task discriminator was designed based on CGAN to generate feature maps for the rendering of NRM accessories.

This paper designs an ACNN with dynamic convolutional kernel characteristics by combining the coordinate attention (CA) (Xuan et al., 2022) and the dynamic field-of-view module. The dynamic field of view module in ACNN is designed using five-layer dynamic cavity convolution. This structure can not only take advantage of dynamic convolution, but also adaptively adjust the feature extraction strategy according to the different input images, so as to better extract the effective features in different wearable scenarios. Firstly, the global average pooling is performed on the original NRL accessory image and the global average under each layer of convolution is calculated as follows:

$$z_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{ijk} \tag{7}$$

where x_{ijk} is the value of the original image at (i, j) and z_k is the average pooling result.

The dynamic field of view module is then utilised to achieve adaptive feature extraction, which relies on the dynamic updating of the dynamic convolutional weights and biases, as shown below:

$$y = g\left(\tilde{W}^T(z)z + \tilde{b}(z)\right) \tag{8}$$

where $\tilde{W}(z)$, $\tilde{b}(z)$ are the weights and bias of the dynamic convolution kernel, respectively.

The difference with static convolution is that dynamic convolution kernel can realise dynamic adjustment of convolution kernel weight and bias according to the different input features in the network, thus improving the expressive ability of the network, which is shown in equation (9).

$$\tilde{W}(z) = \sum_{k=0}^{K} \pi_k(z)\tilde{W}_k, \tilde{b}(z) = \sum_{k=0}^{K} \pi_k(z)\tilde{b}_k \tag{9}$$

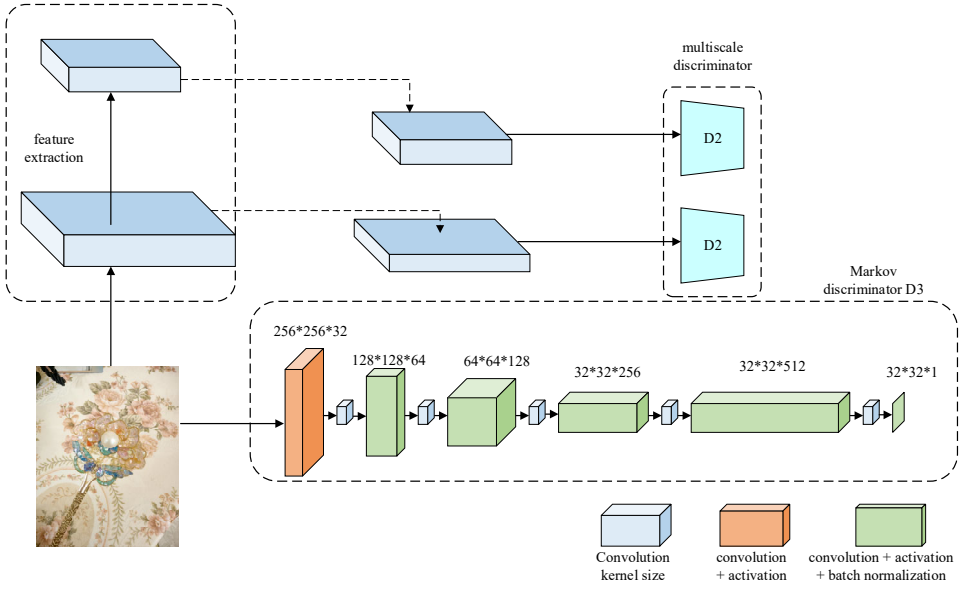
where π_k is the k^{th} attentional weight, $0 \le \pi_k(z) \le 1$, $\sum_{k=0}^{K} \pi_k(z) = 1$, which is dynamically adjusted by the mechanism as the input feature x varies.

According to the method of assigning weights to weights $\tilde{W}_k(z)$ and bias $\tilde{b}_k(z)$ and aggregating them according to π_k respectively, DCNN achieves the ability of dynamic feature extraction for random intangible cultural heritage accessories.

3.2 Multi-task discriminator-based feature map generation for primary non-religious accessories

A multi-task discriminator is used in the feature map generation network to enhance the global and local scale evaluation of the rendering of intangible cultural heritage accessories. The multiscale strategy enables the network to capture different levels of the primary intangible cultural heritage accessory wearable feature maps, so as to evaluate their authenticity in a more comprehensive way. The structure of the multi-task discriminator is shown in Figure 2.

Figure 2 The structure of the multi-task discriminator (see online version for colours)



The multi-task discriminator consists of three key components: D1, D2 and D3, of which D1 and D2 constitute the multi-scale discriminator, which consists of multiple base discriminator units that can analyse and generate feature maps at different scales. In addition, the D3 discriminator adopts the design concept of Markov discriminator (Silva and Narayanan, 2006), which fully considers the local region of the image and can evaluate the local realism of the image in a more detailed way.

The generator consists of a series of residual blocks with upsampled layers, and the multitask discriminator design uses two multiscale discriminators for conditional adversarial loss, a Markov discriminator for two-dimensional cross-entropy loss, and spectral normalisation applied to all convolutional levels. To train the image generator, the total loss function incorporates the conditional adversarial loss, perceptual loss, feature matching loss and 2D cross entropy loss. The expression of the total loss function is as follows:

$$L_{all} = L_{cGAN} + \lambda_\beta L_{FM} + \lambda_\alpha L_p \tag{10}$$

where λ_α and λ_β are different loss weights, L_{cGAN} is the conditional adversarial loss function, L_{FM} is the feature matching loss, L_p and is the two-dimensional cross-entropy loss function.

The multiscale discriminator implementation of conditional adversarial loss uses the hinge loss function (Xu et al., 2017). Hinge loss evaluates model performance by measuring the interval between the model's response to real and generated samples. This loss encourages the discriminator to better distinguish between real and generated samples, which usually leads to more stable training and higher quality generated images. The conditional adversarial loss is calculated as follows:

$$L_{D_R} = -\frac{1}{N}\sum_{i=1}^{N} \min(0, 1 - D(x)) \tag{11}$$

$$L_{D_F} = -\frac{1}{N}\sum_{i=1}^{N} \min(0, 1 + D(x)) \tag{12}$$

$$L_{cGAN} = \lambda_R L_{D_R} + \lambda_F L_{D_F} \tag{13}$$

where L_{D_R} is the sum of the loss functions of the real image, L_{D_F} is the sum of the loss functions of the generated image, N is the number of samples, x is the input localised image region, $D(x)$ is the output of the discriminator, and λ_R and λ_F are the different loss weights.

The feature matching loss technique helps the generator to focus on multiple levels of learning in addition to the final discriminator output, thus obtaining richer and more diverse gradient information. Such a training method usually produces more consistent and higher quality results, as calculated by the following equation:

$$L_{FM}(G, D_k) = E_{(s_x)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left|D_k^{(i)}(s_x) - D_k^{(i)}\left(G\left(s_G, G(s_x)\right)\right)\right|_1\right] \tag{14}$$

where G is the generator, D_k is the k^{th} intermediate level of the discriminator, $E_{(s_x)}$ is averaged over all samples s_x, N_i is the number of features in the i^{th} intermediate level, $D_k^{(i)}(s_x)$ is the feature obtained by s_x through the i^{th} intermediate level in D_k, and $G(s_G, G(s_x))$ is the result of the transformation of G to s_x.

The Markov discriminator uses a two-dimensional cross-entropy loss, which allows the model to focus on each localised region of the image, as shown in equation (15).

$$L_p = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\left(p_{ij}\right) \tag{15}$$

where N is the total number of pixels in the image, C is the total number of categories, y_{ij} is whether pixel i belongs to category j, and p_{ij} is the probability that the model predicts that pixel i belongs to category j.

4 Adaptive perception enhancement-based virtual try-on technology for non-heritage accessories

4.1 Background noise suppression of primary feature maps based on multi-scale feature cascades

Aiming at the issue that the current non-heritage accessories virtual fitting technology is affected by background noise, which leads to poor quality of the generated images, this paper enhances the filtering ability of image background noise by constructing a background noise suppression module, and designs an adaptive perceptual enhancement network structure to reduce the interference caused by the perspective distortion of the images, so as to accurately realise the virtual fitting of non-heritage accessories. As shown in Figure 3, the proposed method consists of two parts: BNR as well as adaptive perception enhancement network.

Figure 3 Adaptive perceptual enhancement-based virtual try-on model for non-religious accessories (see online version for colours)

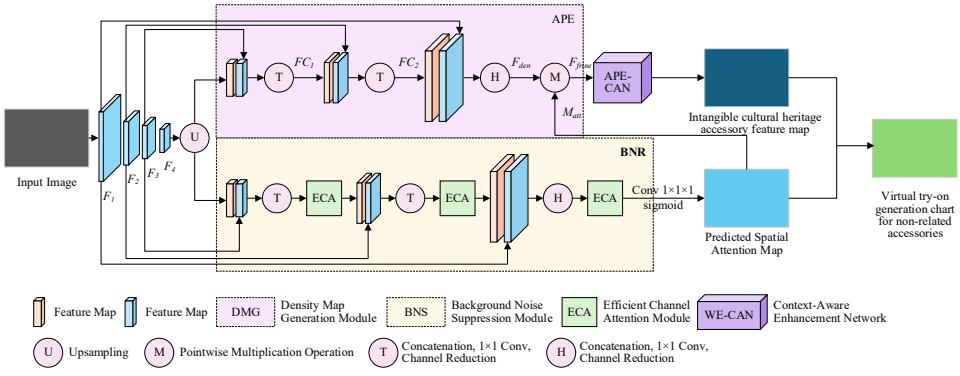
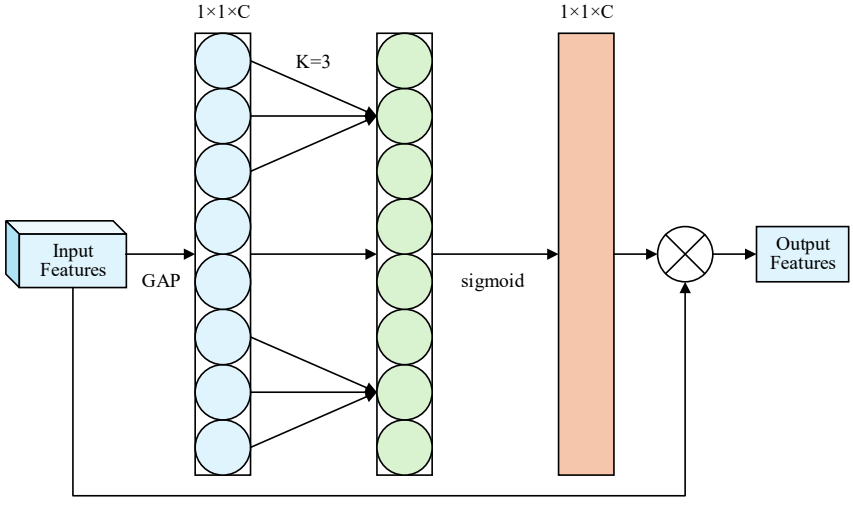


Figure 4 The structure of ECA (see online version for colours)



Through the previous section, this paper obtains the primary intangible cultural heritage accessories try on feature map F_{GAN}, but the feature map has a lot of background noise, which leads to poor image quality, for this reason, this paper designs the BNR module, which is used to suppress the background noise and generate the neural network module for accurate primary intangible cultural heritage accessories try on feature map, as shown in Figure 4. The BNR adopts a design based on ECA modules and multilevel information fusion, aiming to enhance the noise suppression capability of spatial attention maps. M_{att} determines the accuracy of F_{GAN}. To improve the noise suppression of F_{GAN}, BNR takes advantage of the different information of the low-level feature map and the high-level feature map.

ECA is an attentional mechanism used to enhance the expressiveness of CNN features, as implied in Figure 4. It extracts the interaction information between channels from the feature map by using one-dimensional convolution, calculates the weights of different channels, and applies these weights to the feature map. In BNR, ECA utilises local channel attention to enhance the effective information in different features as shown below:

$$y = sigmoid\left(Conv_k\left(GAP\left(F_{GAN}\right)\right)\right) \otimes F_{GAN} \tag{16}$$

where F_{GAN} is the primary feature map; y is the output; sigmoid denotes the activation function; and GAP is global average pooling.

To better highlight the difference between the human try-on parts and the irrelevant background in M_{att}, BNR uses 1 × 1 convolutional dimensionality reduction before using the sigmoid function to map the pixels of M_{att} between (0, 1), thus generating M_{att} that effectively distinguishes the human try-on parts from the irrelevant parts, as shown below:

$$M_{att} = sigmoid\left(W * F_{att} + b\right) \tag{17}$$

where M_{att} is the final generated spatial attention map; F_{att} is the intermediate feature map after 3 ECA optimisations; * is the convolution operation; W and b are the weight and bias of the 1D convolution, respectively.

4.2 Adaptive perceptual enhancement-based virtual try-on result generation for non-religious accessories

To enhance the expression of contextual information and reduce the effect of perspective distortion on the image, this paper designs an adaptive perception enhancement (APE) module, further improves the network structure, and proposes an APE-CAN network, as shown in Figure 5.

APE extracts multi-scale features of F_{GAN} and enhances the expression of its effective information. Firstly, the average pooling of F_{GAN} is carried out, and a 1 × 1 convolution and two 3 × 3 convolution are used to extract multi-scale information of F_{GAN} with different receptor fields. The extracted head features are cascaded and dimensionality reduced after pooling, which is denoted as F_A. The ECA is subsequently utilised to enhance the effective multi-scale information representation in F_{GAN} to generate features with significant information, denoted as F_S. Finally, a multiplication operation is performed on F_A and F_s to generate a multi-scale non-legacy accessory try-on feature map F_m, as shown below:

$$F_m = W * (F_A \otimes F_S) + b \tag{18}$$

Combined with APE to improve the context-aware network (CAN) (Kong et al., 2021), APE-CAN is able to extract more detailed spatial and global information in the features and enhance the network's ability to adaptively optimise the perception of multi-level context information. APE-CAN first performs multi-scale feature extraction on F_{GAN} using four average pooling layers with different kernel sizes to generate four different sized receptive fields to perceive the contextual feature S_j with scales of 1, 2, 3, and 6. Secondly, S_j is subtracted from F_m so as to extract the feature difference C_j between the target feature and the neighbouring features to realise the feature difference extraction, as shown in equation (19). Then, APE-CAN inputs C_j into the weight computation network and computes the scale weights ω_j of different scales in the input feature map using 1D convolution as shown in equation (20). Finally, perspective correction and dimensionality reduction of F_{GAN} using scale weights w_j enables the network to generate a high quality predicted non-legacy accessory virtual try-on image F, as shown in equation (21).

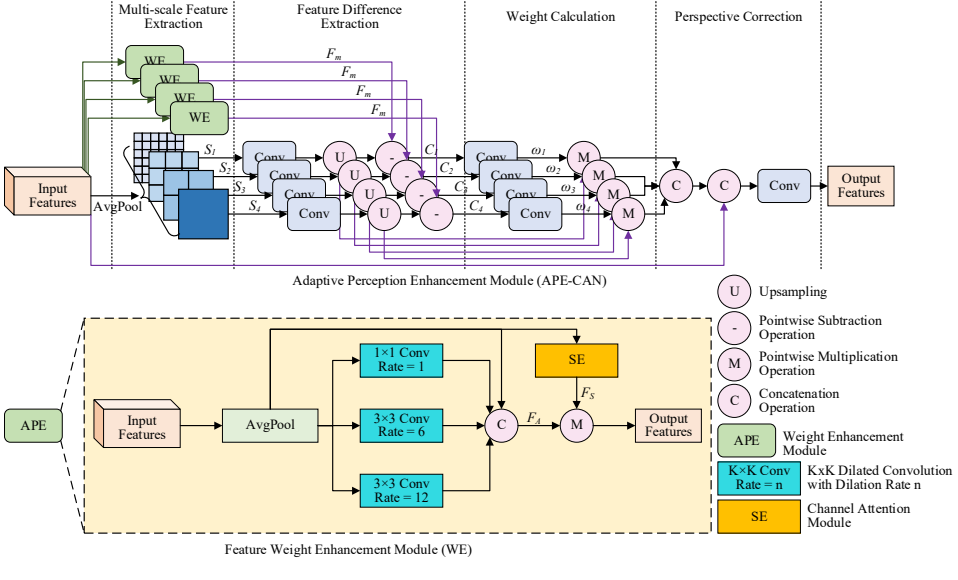
$$C_j = S_j - F_m \tag{19}$$

$$\omega_j = sigmoid\left(W * C_j + b\right) \tag{20}$$

$$F = W * \left((F_{GAN}) \left| \frac{\sum_{j=1}^{4} \omega_j \otimes S_j}{\sum_{j=1}^{4} \omega_j} \right. \right) + b \tag{21}$$

where $(\cdot|\cdot)$is a cascade operation.

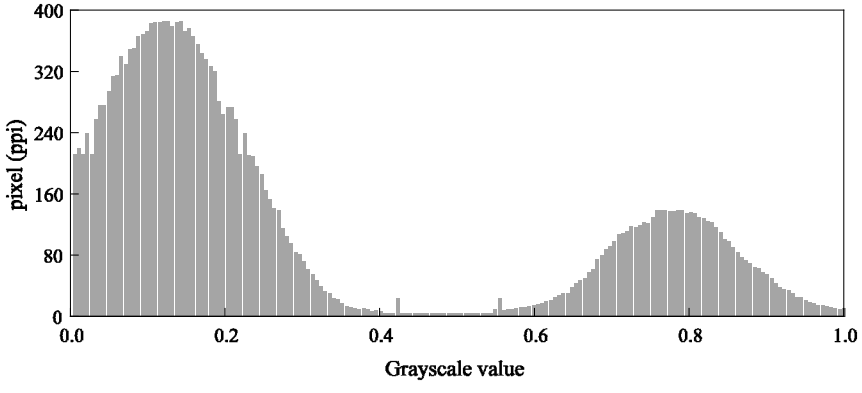
Figure 5 The structure of APE-CAN (see online version for colours)



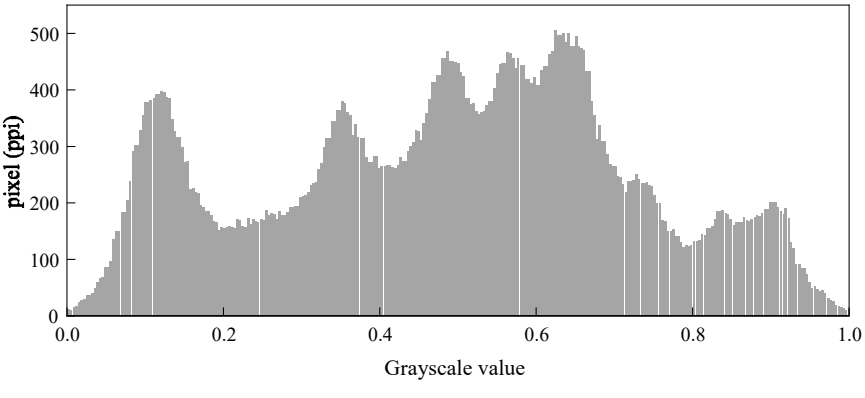
5 Experimental results and analyses

All the experiments in this paper are performed on a computer (PC) with NVIDIA GeForce RTX 1080ti dual graphics cards (12 gigabytes of video memory), and are implemented using PyTorch programming. During training, the Adam optimiser is used, supplemented by learning rate decay. The initial learning rate was set to 0.000 1 and decayed by multiplying the learning rate by 0.98 after every 50th round, for a total of 100 rounds of training. Due to the small number of non-legacy accessory datasets, this paper adopts the UNESCO non-legacy accessory dataset and the VITON virtual fitting dataset as the experimental dataset. UNESCO contains 6,566 pairs of image groups, categorised into target accessories, character images, and before and after depth maps, with a resolution of 512 × 320. The VITON dataset consists of 14,221 sets of matching training images with a resolution of 256 pixels × 192 pixels and 2,032 sets of matching test images with the corresponding resolution. Each set of matching images consists of the foreground half-body image of the human body to be tried on and the corresponding target garment image.

Figure 6 Comparison of original and APE-enhanced image quality on UNESCO dataset, (a) greyscale histogram of original intangible heritage accessory image (b) adaptive perception enhanced image grey level histogram



(a)



(b)

Figure 6(a) shows the grey level histogram of the original NRM accessory image on UNESCO dataset, and Figure 6(b) shows the grey level histogram of the adaptive perceptual enhancement image on UNESCO dataset by the suggested method APE-CAN. The greyscale of the original non-heritage accessories image is mainly concentrated in relatively dark and light areas, using background noise suppression and context-adaptive perceptual enhancement methods to enhance the original image, the greyscale is uniformly distributed, which can effectively improve the effect of non-heritage accessories image enhancement.

To objectively evaluate the generation of virtual fitting images for accessories, this paper uses inception score (IS), SSIM and learned perceptual image patch similarity (LPIPS), peak signal-to-noise ratio (PSNR) for quantitative evaluation of APE-CAN, VGAN (Wang et al., 2022), CAN (Ye et al., 2024), ASNet (Chou et al., 2024) and USAM (Hu et al., 2022) are quantitatively evaluated as shown in Table 1. The IS score is positively correlated with image quality; the higher the score, the clearer and more varied the image. On the UNESCO and VITON datasets, the IS values of APE-CAN are 1.393 and 3.098, respectively, which are at least 1.75% and 3.92% higher compared to the baseline model, respectively. The higher the score of SSIM, the higher the quality of image generation. On the two datasets, the SSIM of PE-CAN was improved by 3.45%–26.76% compared with the other five models. PE-CAN designed a multi-scale discriminator to generate feature maps with multi-scale for virtual try-on of non-heritage accessories, and realised the suppression of background noise of feature maps by SAM, which greatly improved the image generation. The LPIPS score is negatively correlated with the image quality, the lower the score, the higher the similarity between the two images. On both datasets, PE-CAN has the lowest LPIPS score, so PE-CAN is the most effective in generating non-heritage accessory fitting images.

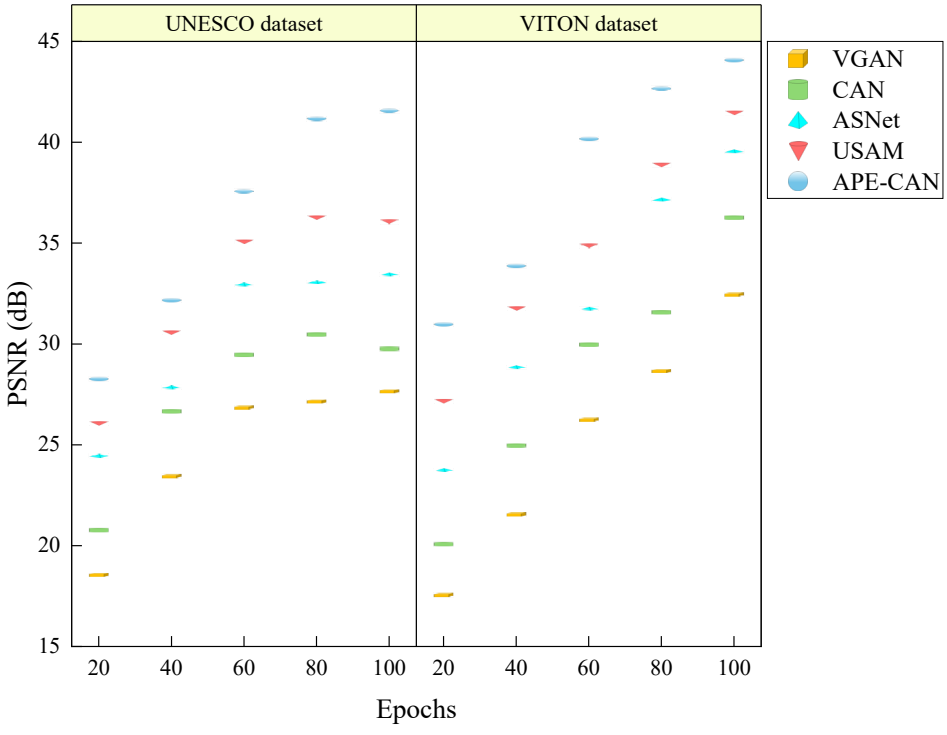
Table 1 Quantitative comparison with mainstream methods

Method	*UNESCO dataset*			*VITON dataset*		
	IS	*SSIM*	*LPIPS*	*IS*	*SSIM*	*LPIPS*
VGAN	1.255	0.686	0.439	1.315	0.791	0.583
CAN	1.321	0.727	0.453	2.498	0.813	0.612
ASNet	1.332	0.741	0.509	2.706	0.854	0.674
USAM	1.369	0.782	0.526	2.981	0.872	0.715
APE-CAN	1.393	0.809	0.575	3.098	0.916	0.739

Figure 7 implies the comparison of PSNR of different methods on both the datasets, higher value of PSNR indicates better image quality with less distortion. After 60 iterations, the PSNR of APE-CAN on UNESCO and VITON datasets reached 37.2 dB and 39.8 dB, respectively, which increased 40.38% and 53.67% compared with VGAN and 27.84% and 34.46% compared with CAN, respectively. Compared with ASNet, the increase was 14.11% and 26.75%, and compared with USAM, the increase was 7.2% and 15.36%. VGAN generates the NRM accessory images only by traditional GAN, but no noise suppression process is applied to the generated images. CAN achieves adaptive semantic perception of images through contextual CNNs, but does not augment the generated images. Although ASNet considers multi-scale feature enhancement, the quality of the pre-generated images of non-heritage accessories is average. USAM performs adaptive perceptual enhancement of the generated images by two-path SAM,

but it does not consider the multi-scale features of the generated images, so the quality of image generation is not as good as APE-CAN. From the above analysis, it can be seen that APE-CAN can generate high-quality virtual try-on images of non-heritage accessories with good performance.

Figure 7 Comparison of PSNR metrics on the two datasets (see online version for colours)



6 Conclusions

Non-legacy accessories carry deep historical and cultural connotations, yet their inheritance and promotion face many challenges, such as the limitations of physical display and the lack of audience experience. In this paper, to address the issue that the existing virtual try-on technology for non-fragrant accessories has insufficient user experience due to the poor quality of image generation, a convolutional neural network that can adaptively adjust the feature extraction strategy according to the non-fragrant accessories try-on scenario is firstly designed, and a multi-task discriminative generative adversarial network is used to generate the primary feature maps to enhance the global and local scale evaluation of the rendering of the accessories. The background noise of the primary feature map is then suppressed based on ECA and multilevel information fusion to remove the noise from the primary feature map. Then feature weight enhancement is introduced to improve the context-aware network, and an adaptive perceptual enhancement module is designed to weight the features at different locations

in the feature map according to the weight information in the feature map to enhance the representation of important features. Finally, the primary feature maps are perspective-corrected and downscaled using scale weights to enable the network to generate high-quality images of non-heritage accessory try-on. Experimental results on real datasets show that the proposed method improves SSIM by at least 3.45% and IS by at least 1.75% compared to the baseline method, with better virtual fitting results.

Acknowledgements

This work is supported by the Joint Science and Education Project of Hunan Natural Science Foundation project named: 'Research on the digital dissemination of Hunan's intangible cultural heritage traditional skills based on virtual simulation' (No. 2022JJ60019).

Declarations

All authors declare that they have no conflicts of interest.

References

Cai, Y., Huo, J., Zhang, H. and Wang, L. (2024) 'Exploration of the innovative development path of bamboo weaving and fashion design in the background of intangible cultural heritage', *Cultural Heritage*, Vol. 6, No. 5, pp.14–23.

Chou, T., Chu, C-H. and Liu, S. (2024) 'Virtual footwear try-on in augmented reality using deep learning models', *Journal of Computing and Information Science in Engineering*, Vol. 24, No. 3, pp.1–10.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A. (2018) 'Generative adversarial networks: an overview', *IEEE Signal Processing Magazine*, Vol. 35, No. 1, pp.53–65.

Dayik, M., Yüksel, H. and Çolak, O. (2016) 'Real-time virtual clothes try-on system', *Development of Polyester Blend Woven Fabric for Better Comfort*, Vol. 10, No. 4, pp.13–18.

Douzas, G. and Bacao, F. (2018) 'Effective data generation for imbalanced learning using conditional generative adversarial networks', *Expert Systems with Applications*, Vol. 91, pp.464–471.

Hauswiesner, S., Straka, M. and Reitmayr, G. (2013) 'Virtual try-on through image-based rendering', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 9, pp.1552–1565.

Hu, X., Zhang, J., Huang, J., Liang, J., Yu, F. and Peng, T. (2022) 'Virtual try-on based on attention U-Net', *The Visual Computer*, Vol. 38, No. 9, pp.3365–3376.

Ishikawa, S. and Ikenaga, T. (2022) 'Image-based virtual try-on system with clothing extraction module that adapts to any posture', *Computers & Graphics*, Vol. 106, pp.161–173.

Islam, T., Miron, A., Liu, X. and Li, Y. (2024) 'Deep learning in virtual try-on: a comprehensive survey', *IEEE Access*, Vol. 12, pp.29475–29502.

Jiang, S., Xu, Y., Li, D. and Fan, R. (2023) 'Self-supervised feature matched virtual try-on', *Journal of Computational Design and Engineering*, Vol. 10, No. 5, pp.1958–1969.

Kong, Y., Feng, M., Li, X., Lu, H., Liu, X. and Yin, B. (2021) 'Spatial context-aware network for salient object detection', *Pattern Recognition*, Vol. 114, p. 107867.

Sabina, O., Elena, S., Emilia, F. and Adrian, S. (2014) 'Virtual fitting–innovative technology for customize clothing design', *Procedia Engineering*, Vol. 69, pp.555–564.

Siddique, N., Paheding, S., Elkin, C.P. and Devabhaktuni, V. (2021) 'U-net and its variants for medical image segmentation: a review of theory and applications', *IEEE Access*, Vol. 9, pp.82031–82057.

Silva, J. and Narayanan, S. (2006) 'Average divergence distance as a statistical discrimination measure for hidden Markov models', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp.890–906.

Tripathi, M. (2021) 'Analysis of convolutional neural network based image classification techniques', *Journal of Innovative Image Processing (JIIP)*, Vol. 3, No. 2, pp.100–117.

Wang, T., Gu, X. and Zhu, J. (2022) 'A flow-based generative network for photo-realistic virtual try-on', *IEEE Access*, Vol. 10, pp.40899–40909.

Wang, Y., Wang, W., Li, Y., Jia, Y., Xu, Y., Ling, Y. and Ma, J. (2024) 'An attention mechanism module with spatial perception and channel information interaction', *Complex & Intelligent Systems*, Vol. 10, No. 4, pp.5427–5444.

Xie, L. and Liao, H.J. (2014) 'A posture recognition method based on skeletal node and geometric relation using Kinect', *Applied Mechanics and Materials*, Vol. 543, pp.2879–2883.

Xu, G., Cao, Z., Hu, B-G. and Principe, J.C. (2017) 'Robust support vector machines based on the rescaled hinge loss function', *Pattern Recognition*, Vol. 63, pp.139–148.

Xu, Q., Liu, H., Liu, Y. and Wu, S. (2021) 'Innovative design of intangible cultural heritage elements in fashion design based on interactive evolutionary computation', *Mathematical Problems in Engineering*, Vol. 20, No. 1, pp.13–16.

Xuan, W., Jian-She, G., Bo-Jie, H., Zong-Shan, W., Hong-Wei, D. and Jie, W. (2022) 'A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection', *IEEE Sensors Journal*, Vol. 22, No. 21, pp.20910–20920.

Ye, J., Wang, Y., Xie, F., Wang, Q., Gu, X. and Wu, Z. (2024) 'Slot-VTON: subject-driven diffusion-based virtual try-on with slot attention', *The Visual Computer*, Vol. 41, pp.1–12.

Zhang, T., Wang, W.Y.C., Cao, L. and Wang, Y. (2019) 'The role of virtual try-on technology in online purchase decision from consumers' aspect', *Internet Research*, Vol. 29, No. 3, pp.529–551.

Zhang, X., Shang, S., Tang, X., Feng, J. and Jiao, L. (2021) 'Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp.1–14.

Zhou, C., Zhang, W. and Lian, Z. (2024) 'Enhancing consistency in virtual try-on: a novel diffusion-based approach', *Image and Vision Computing*, Vol. 148, p.105097.