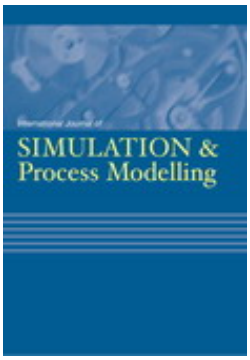
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123
https://www.inderscience.com/ijspm

Simulation modelling of fashion colour harmonisation with visual transformers

Bei Li

DOI: 10.1504/IJSPM.2026.10077921

Article History:

Received:	18 December 2025
Last revised:	26 January 2026
Accepted:	03 March 2026
Published online:	18 June 2026

Simulation modelling of fashion colour harmonisation with visual transformers

Bei Li

Faculty of Art and Design,
Shanghai Business School,
Shanghai, 200235, China
Email: li673612@126.com

Abstract: In the field of fashion design, colour coordination is a critical factor in enhancing market competitiveness. This study aims to develop a simulation model for colour coordination within the fashion design process to improve its quality. First, the simulation model was constructed based on a visual transformer. This model treats the visual transformer as a simulation of the designer's decision-making process. By learning from a large dataset of fashion images, it captures intrinsic patterns between colours and simulates the decision logic designers employ when selecting and coordinating colours. As a workflow simulation model, it focuses not only on colour coordination within images but also on emulating the fashion design workflow itself. Simulation experiments conducted on the dataset demonstrate that the proposed model achieves an image quality distance of 5.09 and a colour richness of 40.14, outperforming the comparison model. Significant improvements are observed in colour harmony and image generation quality.

Keywords: fashion design; image colour coordination; visual transformer; ViT; process simulation; workflow modelling.

Reference to this paper should be made as follows: Li, B. (2026) 'Simulation modelling of fashion colour harmonisation with visual transformers', *Int. J. Simulation and Process Modelling*, Vol. 23, No. 2, pp.77–89.

Biographical notes: Bei Li is a Lecturer at the Faculty of Art and Design, Shanghai Business School, and a member of the China Fashion Association, China. She obtained her Bachelor's degree (2009) and Master's degree (2012) from Donghua University, China. During this period, she also earned a Master's Degree from Istituto Europeo di Design (Milan, Italy) in 2011. She has published four papers. Her research interests include fashion design, image colour matching, and visual transformer.

1 Introduction

Colour, as a core element of fashion design, serves not only as the primary vehicle for visual aesthetics but also as a crucial medium for conveying brand concepts and shaping product styles (Yum, 2023). In traditional fashion design workflows, the colour coordination process heavily relies on designers' experiential intuition and subjective judgement, lacking systematic scientific theoretical support (Yan et al., 2023a). As the fashion industry enters a phase of digital transformation, integrating artificial intelligence technology with design process simulation to construct efficient colour harmony simulation models has become a core direction for achieving design automation.

Transformer models, with their self-attention mechanism and global feature modelling capabilities, have revolutionised computer vision and established a new paradigm for complex visual semantic understanding (Xu, 2025). Visual transformer (ViT) precisely extracts global semantic features and local detail correlations by dividing images into discrete visual tokens and leveraging self-attention to capture long-range dependencies between tokens (Mrinali and Gupta, 2025). Its modelling logic aligns closely with designers' colour selection decision-making processes. This paper constructs ViT as a core simulation model that mimics designers' colour decision-making processes. By introducing it into the field of apparel colour coordination design, it overcomes the limitations of traditional methods focused solely on image processing. This achieves collaborative simulation modelling for colour harmonisation, providing scientific and technological support for process simulation in fashion design colour coordination (Gu et al., 2023).

The central challenge in clothing colour coordination is modelling colour compatibility and stylistic coherence. Early research primarily employed rule-based heuristic methods. The early rule-based approach to clothing colour coordination centres on classical colour theory and industry expertise, breaking down colour pairing logic into fixed, actionable rules. These rules generate matching schemes through matching and filtering processes. However, this traditional method focuses solely on quantifying colours through objective numerical values, neglecting the physical properties of garments that influence colour perception. Consequently, colour combinations that appear numerically

sound often result in visual dissonance when worn. Dong et al. (2023) constructed clothing feature vectors by extracting colour histograms and texture features, then measured item similarity using Euclidean distance. However, their manually selected features struggled to capture higher-order abstract styles. Hu et al. (2019) pioneered the integration of collaborative filtering into outfit recommendations by constructing co-occurrence matrices from users' historical colour pairing data. This algorithm relies on historical user colour combination data to construct a co-occurrence matrix. However, in real-world scenarios, the proportion of clothing styles purchased by users is extremely low, resulting in an extremely sparse user-clothing rating matrix. The cold-start problem caused a sharp decline in recommendation effectiveness for new items. Shamoi et al. (2020) treated clothing colours as time-series data to learn pairing patterns, achieving a matching accuracy of 63.4% on the Polyvore dataset. However, their fixed-order assumption limited flexibility in cross-category combinations. Sang (2025) pioneered the application of decision trees, modelling attributes such as colour, category, and material as tree nodes. By learning compatibility propagation between nodes through an attention mechanism, the matching accuracy was improved to 71.2%. Huang (2020) employed contrastive learning to map textual descriptions and colour features into a shared space, enabling zero-shot generation from text to outfits. However, purely vision-driven models struggle to interpret users' implicit semantic intentions.

Early studies primarily relied on qualitative descriptions such as soft and harmonious or strong contrast, lacking quantifiable metrics. This led designers to depend on subjective interpretations, making precise colour coordination difficult to achieve. As the deep learning is rapidly growing, deep neural networks have made remarkable strides in image processing through their powerful feature learning capabilities, gradually achieving superior colour coordination results. A deep learning-based clothing colour coordination model overcomes the limitations of traditional rule-based methods' static logical constraints and collaborative filtering's data dependency bottlenecks. By performing deep feature extraction, complex association modelling, and dynamic pattern learning on massive colour combination datasets, it intelligently enhances colour harmony across dimensions including the visual characteristics of colours themselves, the aesthetic logic of combinations, and the compatibility between scenes and subjects. First, colour harmony hinges on accurately capturing visual characteristics and associative properties of colour combinations. Deep learning models leverage deep neural network architectures to extract multidimensional features from colours themselves, their carriers, and contextual pairings, establishing a foundation for harmony assessment. Second, the deep learning model eliminates the need for manually defined colour pairing rules. Instead, it directly utilises vast high-quality clothing colour combination samples as training data. Through forward and backward propagation within the network, it automatically uncovers nonlinear relationships among multiple colours, thereby enhancing the harmony of colour pairings. Li et al. (2025) proposed a clothing colour coordination method based on clustering and convolutional neural networks (CNNs). Given a greyscale garment image as input, the method automatically searches for the closest cluster, extracts each pixel as input for the CNN, and finally employs a combined bilateral filter to further optimise the output colour image. He et al. (2025) proposed a dual-stream network-based garment colour coordination model. This model utilises an image colourisation method that combines global and local features. Cao et al. (2025) extracted local and global features from each layer of greyscale clothing images, performed feature fusion, and established a multi-layer feature extraction network, thereby improving image colourisation results to some extent. CNN-based clothing colour coordination methods face significant challenges in obtaining paired image datasets containing both colour and greyscale versions for training neural network models, which severely limit the application of CNNs. The feature extraction of CNN has extremely high requirements for the quality of the input image. If the input image has issues such as low resolution, uneven lighting, colour deviation, occlusion, blurriness, etc., which are noise and distortions, the local feature extraction will suffer from severe distortion, ultimately leading to completely incorrect judgements on colour coordination and the results of scheme generation. In the actual scenarios of fashion design, the input of designers may be hand-drawn sketches, low-resolution reference images, or poorly lit physical photography images. Such non-standardised inputs will significantly reduce the performance of the CNN model and even cause it to fail completely. Yan et al. (2023b) proposed a model using generative adversarial networks (GANs) originally designed for converting greyscale clothing images to colour ones. Their generator network adopted a U-Net architecture, enhancing the harmonisation of colour combinations. Zhao et al. (2023) discussed a colour coordination method using encoded data. However, this approach neglected the capture of minute details, resulting in significant discrepancies compared to real images. In recent years, the introduction of ViT has revitalised traditional greyscale image colouring models. Compared to conventional neural network models, ViT demonstrates superior global feature modelling capabilities. Zhang et al. (2023) employed a ViT model to design a point-interactive clothing colour coordination method, though it exhibits high computational complexity. Zhao et al. (2024) leveraged the global receptive field of ViT's self-attention layer, enabling the model to selectively propagate colour semantics to relevant regions within each individual layer. Concurrently, they employed local stable layers for an efficient upscaling process. Zou et al. (2025) proposed an end-to-end ViT-based clothing colour coordination model, further enhancing the semantic consistency and colour richness of the colouring results.

Through a comprehensive analysis of existing clothing colour coordination methods, we identify a critical limitation: current approaches often lose significant information during image processing. Capturing long-range colour dependencies typically requires expanding convolutional kernels or increasing the receptive field through multiple layers, both of which incur substantial computational costs. To address the aforementioned issues, this paper proposes a ViT-based simulation modelling approach for fashion colour coordination. By deeply embedding ViT into a simulation framework that integrates human colour perception with designers' colour-matching decisions, the core innovation lies in reconstructing apparel colour coordination technology through a simulation modelling paradigm. This transforms simple colour matching into a standardised simulation within the fashion design process, manifested in the following four aspects.

1 A novel encoding-decoding framework with enhanced feature learning. This paper introduces an integrated encoding-decoding architecture that fundamentally improves feature representation. The encoder employs spatial normalisation and feature matching to learn consistent and expressive colour features. The decoder utilises a spatial normalisation residual network to inject information into quantised features. This co-design enhances output diversity while preventing artefacts and guides the encoder toward more robust feature representations, significantly strengthening the model's ability to understand and represent colours in clothing images.

2 A unified ViT-based architecture for holistic colour understanding. Proposing a dedicated ViT model for apparel colour coordination. Its core innovation lies in unifying local detail extraction and global dependency modelling within a single, efficient architecture. By leveraging the self-attention mechanism, the model directly captures long-range colour relationships across a garment, eliminating the need for computationally expensive deep convolutional stacks. Furthermore, this architecture is inherently adaptable, enabling effective training on similar visual tasks without structural changes, thereby improving overall performance and flexibility.

3 An efficient transformer block has been designed. The window-based multi-head self-attention mechanism significantly reduces computational complexity. ResNet mitigates noise interference in image processing and prevents overfitting caused by excessive network depth. The locally augmented forward propagation network enhances the model's ability to extract and utilise local information.

4 Simulation experiments on colour harmony were conducted using public datasets. Results indicate that the proposed model achieved at least 11.66% and 5.04% improvements in colour fidelity (CF) and structural similarity (SSIM), respectively, compared to the baseline model. This research provides an intelligent colour coordination solution for apparel design, effectively lowering design barriers while enhancing design efficiency and aesthetic quality of works. It holds significant implications for advancing the digital design transformation within the apparel industry.

2 Relevant theoretical analysis

2.1 Visual transformer

Transformer is a complex algorithmic model centred on the multi-head attention mechanism (Guerra and Mota, 2006). Given its exceptional performance in processing long sequence data, effectively applying transformers to two-dimensional image data has become a significant research focus. It is against this backdrop that ViT emerged. The innovation of ViT lies in treating images as a special type of sequence (Li et al., 2024). Specifically, ViT employs image patch preprocessing techniques to divide images into multiple sub-patches of fixed size. Each patch is treated as a token within a sequence. These patches are then modelled through an auto-attention mechanism. This approach not only preserves the global modelling capabilities of transformers but also overcomes the limitations of traditional CNNs in modelling complex scenes. ViT consists of the following components.

1 Image patching: ViT discretises the input image into multiple equal-sized two-dimensional image patches, with each patch treated as a token. These patches are transformed into a high-dimensional embedding space via a learnable linear projection matrix, forming patch embeddings.

2 Position encoding: ViT employs a position encoding injection module to embed additional position-dependent information for each image patch, thereby explicitly modelling the image's spatial structure.

3 Transformer encoder: in ViT's feature learning process, multi-scale attention (MSA) first performs global context aggregation, dynamically modelling semantic relationships between image patches via attention weights. Subsequently, a feedforward neural network (FFN) refines the aggregated features, introducing nonlinear expressive power to capture complex visual patterns.

4 Classification head: ViT employs a dedicated classification token to aggregate global information. Finally, the classification token's features are mapped to the category space via a fully connected layer, completing the image classification task (Zhang and Yang, 2021).

The ViT overcomes the limitations of CNNs' local inductive bias and RNNs' sequential computation bottlenecks through global self-attention modelling, parallel

computation, and multi-scale feature fusion, enabling more efficient image processing. In the field of apparel colour coordination, its global semantic understanding and cross-modal adaptation capabilities position it as a core technology driving intelligent fashion design. By processing multiple attention heads in parallel, ViT simultaneously captures local details and global dependencies, avoiding the loss of holistic information inherent in CNN's layer-by-layer stacking approach. Furthermore, its self-attention mechanism enables explicit learning of long-range dependencies, resolving the information decay in long sequences caused by RNN's sequential computation and significantly enhancing the efficiency of temporal image processing.

2.2 Residual neural network

Traditional CNNs suffer from feature degradation when network depth increases, as features cannot propagate backward effectively. Accuracy gradually saturates and declines, leading to network degradation. Residual neural networks (ResNet) first introduced residual connections to address this issue, optimising network performance (Xu et al., 2023). The residual connection module combines original input features with pre-processed features, enabling subsequent layers to utilise all detailed features contained in the original input. This module also ensures the feature extraction capability of deep networks by enabling the identity mapping of shallow features to deep layers. It allows shallow features containing more detailed information to be propagated to deeper layers, facilitating cross-layer feature flow. Simultaneously, during backpropagation, gradients can propagate from deep to shallow layers, preventing gradient vanishing and enhancing network performance (Cao, 2025).

Traditional CNNs use identity mapping. Identity mappings degrade into accuracy of a shallow network when used in deep networks; therefore residual connections transform identity mapping function $H(x) = x$ into an identity mapping function with added residuals $H(x) = F(x) + x$. Consequently, the fitting function that needs to be learned by the network becomes $F(x) = H(x) – x$. As long as $F(x) = 0$, it can revert back to identity mapping function $H(x) = x$.

ResNet significantly enhances CNN performance through key optimisations such as introducing residual connections and batch normalisation, achieving breakthroughs particularly in addressing the challenges of training deep networks. However, these optimisations do not entirely eliminate the inherent limitations of CNNs. Rather, they provide models with greater robustness and generalisation capabilities by mitigating specific issues.

3 Clothing image encoding based on spatial normalisation and feature matching

3.1 Spatially normalised encoder and decoder

Garment image encoding is a key step in transforming visual information into computable, analysable mathematical representations, providing the basis for deep learning models to understand the complex rules of colour coordination. Since encoders cause partial loss of colour information when encoding garment images, features extracted from similar colour blocks tend to have high similarity. Additionally, vector quantisation employs a similarity measurement method, which often maps similar colour regions into the same discrete code index, resulting in repeated index selections and subsequently generating similar image patches after decoding. Furthermore, the current mainstream methods for encoding clothing images all encounter core problems such as distortion in extracting colour information due to the mismatch between the general framework and the specific requirements of clothing colour coordination. During the vectorisation process, repeated index selection will further exacerbate the loss of feature information, the failure of vector representation, and even introduce irreversible cumulative errors, becoming a hidden technical bottleneck for clothing colour coordination.

To address the issue where existing garment image encoding methods cannot restore genuine colour information from garment images, this paper proposes a garment image encoding method based on spatial normalisation and feature matching. This method uses a spatially normalised residual network in the decoder to inject information into quantised features, enhancing diversity generation and avoiding artefacts. At the same time, it introduces feature matching into the discriminator to guide the encoder in learning more consistent and representative feature representations, improving the model's ability to understand and represent colours in garment images. This paper adopts the concept of spatial normalisation (SPADE) (Park et al., 2019) to activate encoded features after discretisation, mapping similar colour regions into the same discrete code index. After spatial transformation, diverse features can still be decoded. The core mechanism of the SPADE involves adjusting normalised activation values using semantic segmentation maps or other spatial guidance information to achieve precise representation of image semantic features. In this study, the SPADE core mechanism activates discretised encoded features. By mapping similar colour regions to the same discrete code index and implementing spatial transformations, it achieves diversified feature decoding. This process precisely simulates the cognitive logic of designers categorising similar colours while preserving regionally differentiated features. SPADE performs denormalisation on normalised activation values using external data, following these specific implementation steps.

Quantised feature maps are first projected into an embedding space via convolution, and then parameters β and γ are learned through two convolutional layers. The

activation value at position (c, y, x) of the previous layer is as follows. The coordinate parameters (c, y, x) of the feature map represent the channel, height and width dimensions respectively, $h_{c,y,x}$ represents the original feature value of the feature map at coordinate (c, y, x) before being activated by SPADE, and h represents the final activated feature value obtained after being adjusted by SPADE. The mapping relationship between them is established through the following equation:

$$h = \gamma_{c,y,x} \frac{h_{c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x} \tag{1}$$

where $h_{c,y,x}$ represents the feature value before activation, μ_c and σ_c are the mean and standard deviation of this channel, h is the activated feature value at coordinate (c, y, x) after being adjusted by SPADE, μ_c and σ_c respectively represent the mean and standard deviation of all feature values on the c^{th} channel, which are used for basic normalisation calculation. $\gamma_{c,y,x}$ and $\beta_{c,y,x}$ are the coordinate-adaptive adjustment parameters learned by SPADE through spatial guidance information, aiming to avoid homogenisation of similar colour regions' features. This characteristic precisely meets the requirements of clothing colour coding, ensuring consistency in colour representation across channels for similar colours, and also distinguishing subtle colour differences at different spatial positions.

The encoder in this paper consists of a sequence of convolutional blocks, residual blocks, downsampling modules, and attention modules. The input image is 256 × 256 × 3, which is compressed to 16 × 16 × 1,024 using an upsampling factor of 16. For the decoder structure, SPADE and U-Net modules are integrated before sampling in the original decoder to inject information into the quantised feature map.

3.2 Feature-based discriminator

The discriminator is a key component of generative adversarial networks, and its role is to evaluate the authenticity of generated images or data. This paper uses the PatchGAN (Chen et al., 2023) model. Unlike traditional GANs, PatchGAN does not generate entire images but instead discriminates local regions (patches) within images. Traditional GANs consist of several convolutional layers and a fully connected layer, extracting image features through convolutions and then outputting a single value via the fully connected layer, which represents the probability that the discriminator considers the input data to be real. PatchGAN is a variant of the generative adversarial network designed for fine-grained discrimination of local features. It differs from the traditional GAN's global image-level discrimination logic and instead adopts local patch-level discrimination, making it more suitable for visual tasks related to clothing colour coordination. When PatchGAN is integrated into the clothing colour coordination model, it can fundamentally solve the problems of global colour imbalance and local colour discontinuity that often occur in traditional GANs during colour coordination, thereby significantly improving the colour coordination effect. PatchGAN is designed as a fully convolutional model; after passing through various convolutional layers, the input does not go into a fully connected layer or an activation function but instead uses convolutions to map the input into an $N \times N$ matrix. Each point in this matrix represents the evaluation of a small region in the original image.

To further improve the performance of the discriminator, shallow feature matching is implemented within the discriminator in this paper. The goal of feature matching is to measure similarity between generated and real data by comparing features extracted from intermediate layers of the discriminator; this helps the discriminator better capture subtle characteristics and high-level structures of the data. Both original clothing images and generated clothing images are fed into the discriminator, while computing shallow feature-matching losses for the first three layers that more easily capture global colour features in images. For the last two layers, adversarial loss from PatchGAN is directly used as constraints. This method enables the model to extract and match features from real and fake images at different scales, which helps learn richer colour information and more accurately recover the structure of the original image. The loss for layer t is as follows, where t represents the number of layers in the discriminator, N denotes the number of features in layer t of the discriminator, x refers to the input image, x' is the reconstructed image, and $Dis_t^i(x)$ and $Dis_t^i(x')$ represent the i^{th} feature values extracted from the original and reconstructed images, respectively, at layer t of the discriminator.

$$L_{fm}(t) = \frac{1}{N} \sum_{i=0}^{N} \left[\left| Dis_t^i(x) - Dis_t^i(x') \right| \right] \tag{2}$$

For the PatchGAN discriminator used for colour coordination of clothing, its core task is to identify the feature deviations of the generated clothing colours on local patches. The shallow features retain the most original local visual details, and the feature matching enables the discriminator to perceive the subtle feature differences at the pixel level/local patch level, rather than merely being able to recognise the obvious deviations at the global level, thereby solving the problem that traditional discriminators are insensitive to fine-grained colour deviations.

3.3 Vector quantisation process

In the vectorisation process for clothing images, first a codebook is learned using the clothing encoder, named $V_{garment}$. By means of this codebook, arbitrary image information x can be discretised. To convert continuous image content into discrete sequences, a clothing encoder E and a clothing decoder D are first learned. The encoder maps the image to an intermediate vector Z via convolution; for each vector at position (i, j) in Z, a quantiser $q(\cdot)$finds its nearest neighbour e_k in the codebook using the nearest-neighbour algorithm, and replaces it, where k is the index. The whole process is as follows:

$$\hat{Z} = q(Z) = \left(\arg\min_k \left\| Z^{i,j} - e_k \right\|\right) \tag{3}$$

After the above steps, the original image has been discretised into a collection of several vectors from the codebook. Then, the quantised feature map $\hat{Z}$ is input to the decoder D to obtain the reconstructed clothing image, as follows:

$$x' = D\left(\hat{Z}\right) = D(q(Z)) \tag{4}$$

Since the quantisation process uses the nearest-neighbour algorithm to replace the original features, this process is non-differentiable, leading to the encoder being unable to compute gradients and train; therefore, the decoder gradient is directly copied to the encoder. The loss function of the entire process is as follows:

$$L_{VQ} = \|x - x'\|^2 + \left\|sg[E(x)] - \hat{Z}\right\|_2^2 + \left\|sg\left[\hat{Z}\right] - E(x)\right\|_2^2 \tag{5}$$

where the vector quantisation loss function L_{VQ} is a composite of three loss terms, designed to comprehensively constrain the differences between the original image and the reconstructed image, as well as the alignment between encoded features and quantised features, $sg[\cdot]$performs gradient stopping, whose core role is to halt gradient backpropagation, ensuring training convergence.

4 Visual transformer-based clothing colour coordination model

4.1 Network framework

To address the issue that current models neglect local details and global information in clothing colour coordination, this paper proposes a clothing colour-matching model based on ViT (UNViT). However, since transformer networks generally consume too many resources, to reduce computational costs, a window-based multi-head self-attention mechanism is introduced. In terms of image information fusion, to better pass the lost image information during downsampling into the upsampling process, two skip connection methods are used to transmit this lost information for subsequent computations. In addition, since images often contain a lot of task-irrelevant noise, which may more or less affect subsequent colouring operations, UNViT is introduced here to suppress noise in the image. At the same time, due to the inherent characteristics of residual networks, overfitting caused by excessively deep networks is prevented.

Figure 1 The overall architecture of the UNViT model for clothing colour harmonisation (see online version for colours)

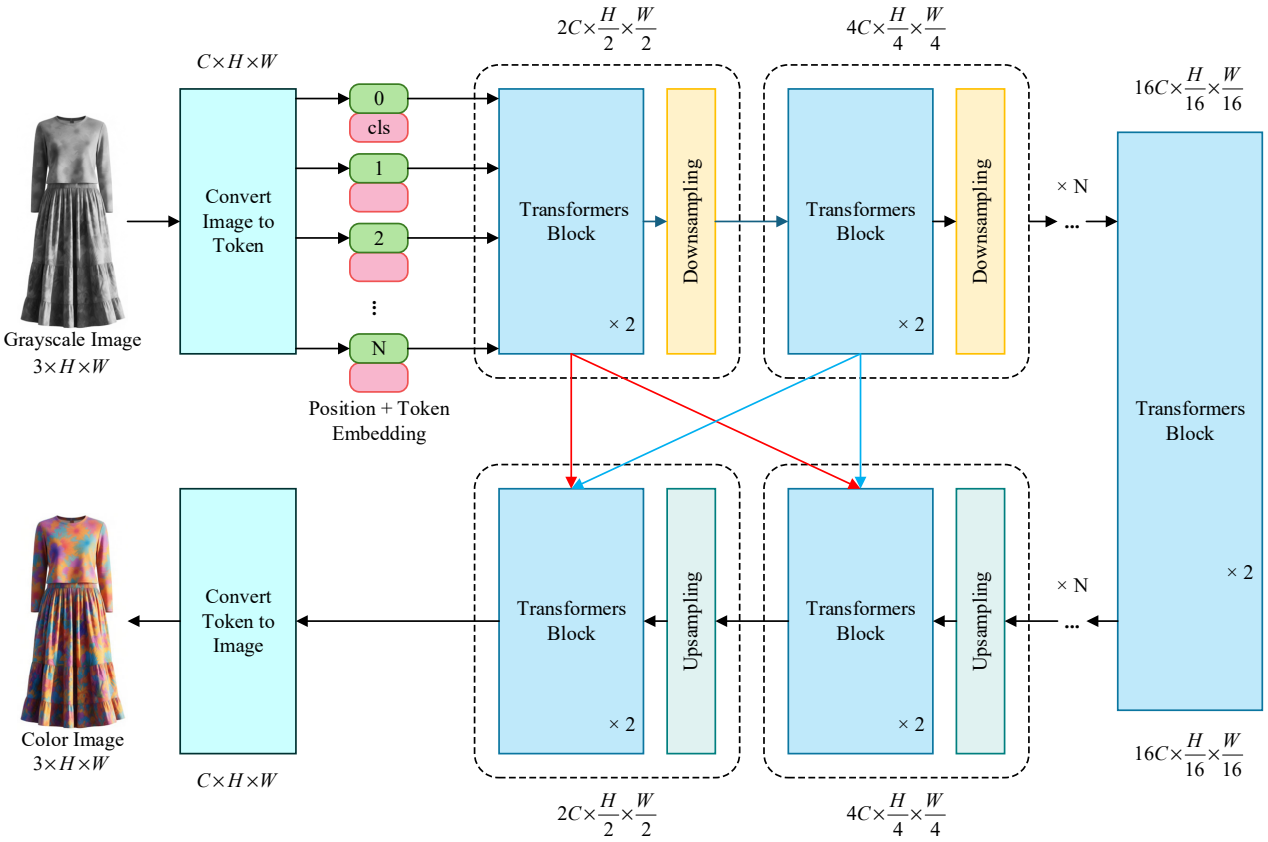


Figure 1 shows the overall architecture of the UNViT model constructed in this chapter. The general structure of UNViT is very similar to that of the classical U-Net network in shape, and the internal part of the model consists of many specially designed transformer blocks. The greyscale clothing image to be matched and the colour-matched image after pairing are respectively used as the input and output of the model. In the process of image colour coordination, to promote feature fusion and usage of images, the red-blue connection line in Figure 1 represents that dual skip connections are used to connect the transformer blocks of the encoder and decoder.

The clothing image inputted into ResViT is a three-channel greyscale image $I \in R^{3\times H\times W}$. First, a convolutional layer with a LeakyReLU activation function 3 × 3 is used to extract shallow features from the image, thus obtaining feature maps $X_0 \in R^{C\times H\times W}$. Then following the design concept of *U*-shaped network architecture, X_0 goes through l encoder combination blocks. Each encoder combination block contains two stacked transformer blocks and one downsampling layer. The downsampling layer consists of a convolutional layer with kernel size 4 and stride 2. After passing through each downsampling layer once, the size of the feature map will be reduced by half while its number of channels doubles, X_0 after going through l encoder combination blocks, the size of the feature map becomes $X_l \in R^{2^l C\times \frac{H}{2^l}\times \frac{W}{2^l}}$.

After passing through the l encoder combination blocks, two transformer blocks are added as the bottleneck stage. In U-Net, the main role of the bottleneck stage is to fuse the deep features extracted from the downsampling part with the shallow features from the upsampling part, allowing the model to switch between high and low resolutions, thus ensuring both high resolution and effective clothing colour coordination.

Then comes the decoder stage. Similar to the encoder stage, the decoder stage consists of l decoder combination blocks stacked together, and each block is composed of two transformer blocks and an upsampling layer. The upsampling layer uses a 2 × 2 transposed convolution with stride 2 for upsampling. After passing through the upsampling layer, the number of feature channels is halved while the size of the feature map doubles. Then, the upsampled features are concatenated with corresponding features from the encoder via skip connections. After going through l decoder combination blocks, finally a 3 × 3 convolutional layer converts the feature map with size $X_l \in R^{2^l C\times \frac{H}{2^l}\times \frac{W}{2^l}}$ into a three-channel colour image with size $R \in R^{3\times H\times W}$, completing the clothing image colour coordination process.

4.2 Enhanced U-Net network

In U-Net, skip connections are typically used to concatenate the information lost during downsampling in the encoder stage with the upsampling features from the decoder stage for feature fusion. This allows the decoder to retain more high-level feature details when performing upsampling and thus restore colour information in the original clothing image more accurately and completely, improving the model's colour coordination capability. However, U-Net usually directly performs cross-layer feature reconstruction between the current encoder layer and the corresponding decoder layer (Williams et al., 2023). Let x be the input features; $downsampl(x)$ represents downsampling operations in U-Net, while $upsample(x)$ represents upsampling operations, where + indicates element-wise addition operation. Skip connection can be represented as follows:

$$skip(x) = downsample(x) + upsample(x) \tag{6}$$

This paper introduces cross-layer and cross-neighbouring layer feature reconstruction on the basis of U-Net to improve the overall performance of the network. This method not only enhances information transmission between layers but also makes full use of detailed information in high-level convolutional feature layers, thereby maximising the overall performance of the network. At this time, skip connections change into equation (7), where x is the feature from the current downsampling layer, and x_1 and x_2 are features from different upsampling layers.

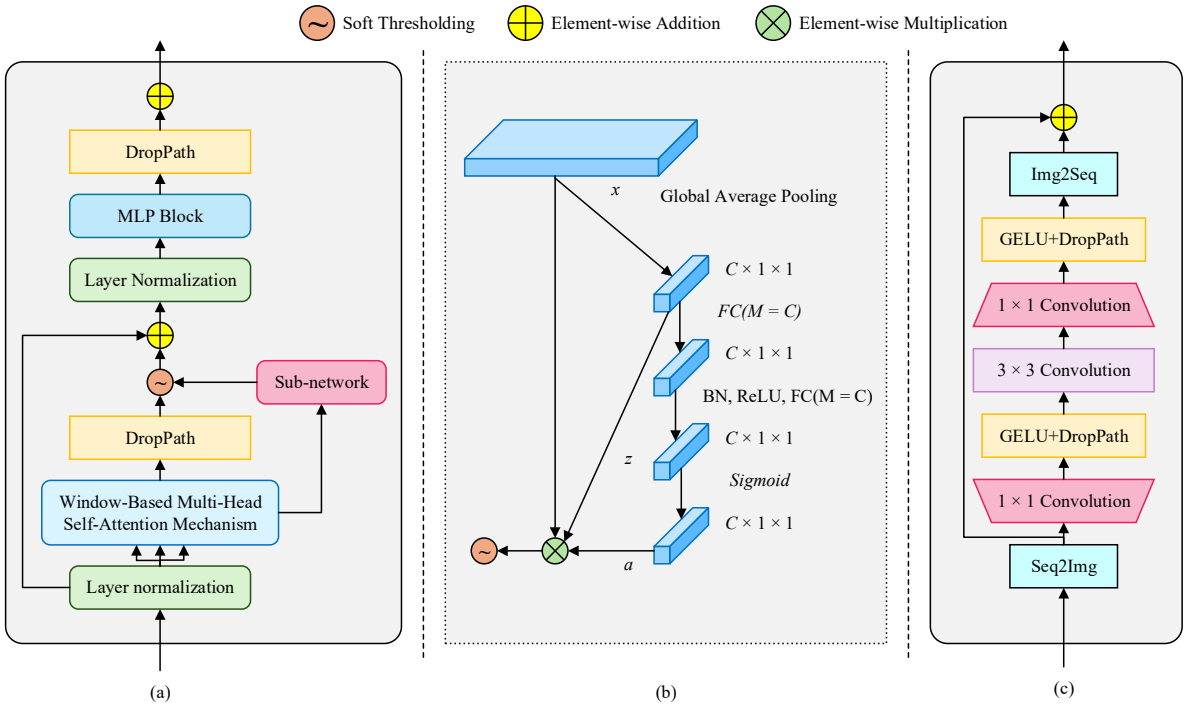
$$\begin{aligned} skip(x) = {} & downsample(x) + upsample(x_1) \\ & + upsample(x_2) \end{aligned} \tag{7}$$

Dual skip connections enhance feature representation capability through cross-same-layer and cross-adjacent-layer feature reconstruction. This method improves the quality of feature representations by reconstructing more layer features during training.

4.3 Design of the transformer block

Using ViT for clothing image colour coordination mainly considers two aspects. First, the classic ViT architecture globally computes self-attention mechanisms between all tokens; every time the number of tokens doubles, the computational load increases quadratically. Second, for images, local feature information is very important because these details help describe the image at a more refined level and provide valuable insights for colour coordination in images. Researchers have found that ViT is better suited to capturing global dependencies in images but has limitations when it comes to capturing local dependencies. In contrast, CNNs are particularly good at extracting local features from images, so combining them can lead to improved performance in image colour coordination.

Figure 2 The window-based transformer block with a subnetwork (see online version for colours)



As shown in Figure 2, this paper designs a window-based transformer block with a subnetwork, consisting of layer normalisation (LN), window-based multi-head self-attention (WMSA), a subnetwork, DropPath regularisation, and a forward propagation network LeFF with local enhancement provided by the multilayer perceptron. The entire computational process for the designed transformer block is as follows:

$$\begin{cases} X' = LN(X) \\ X'' = DropPath\left(WMSA\left(X'\right)\right) \\ \quad \ominus Sub\text{-}Network\left(W-MSA\left(X'\right)\right) \\ X''' = X'' \oplus X' \\ X'''' = DropPath\left(MLP\left(LayerNorm\left(X'''\right)\right)\right) \\ X_{final} = X''' \oplus X'''' \end{cases} \tag{8}$$

where X is the input. X is fed into the LN layer to obtain X'. Next, X', processed by WMSA, is fed into both the DropPath regularisation and a sub-network, and then undergoes soft-thresholding $\ominus$ to yield X''. Subsequently, X' and X'' are connected via a residual connection $\oplus$ to produce X'''. Then, X''' is processed sequentially by an LN layer, an MLP block, and DropPath regularisation, resulting in X''''. Finally, X'''' and X''' are connected via a residual connection to obtain the final output X_{final}.

LN normalises the activation of layers in deep neural networks by subtracting the mean and dividing by the standard deviation of the activations, which helps reduce internal covariate shift and improve convergence during training, where the computation for LN is as follows:

$$LN(X) = \frac{X-\mu}{\delta} \circ \gamma + \beta \tag{9}$$

The object acted upon by the LN layer is $X \in R^{d_k}$, with $\mu \in R$ and $\delta \in R$ being the mean and variance respectively for each sample, $\gamma \in R^{d_k}$ a learnable scale parameter, $\beta \in R^{d_k}$ a learnable shift parameter, of dimensionality d_k.

Dropout and DropPath are both classic regularisation techniques in deep learning. Their core objective is to alleviate overfitting by randomly discarding some network nodes. However, there are fundamental differences in the objects being discarded, the dimensions of their effects, the implementation logic, and the compatibility with network architectures between the two. Dropout focuses on randomly discarding neurons within a single layer and is a general regularisation method applicable to fully connected layers. DropPath, on the other hand, focuses on randomly discarding the entire residual branch in the network and is a regularisation method specifically designed for deep residual networks. The DropPath mentioned in this paper shares similar characteristics with dropout. However, their difference lies in the fact that DropPath randomly samples multiple sub-paths for inactivation, while dropout achieves its function by randomly removing connections between neurons; thus, the scope of DropPath is broader than that of dropout.

4.4 *Window-based multi-head self-attention mechanism*

ViT's computational complexity is in a quadratic relationship with the number of tokens. In computer vision, even processing clothing images at low resolution can result

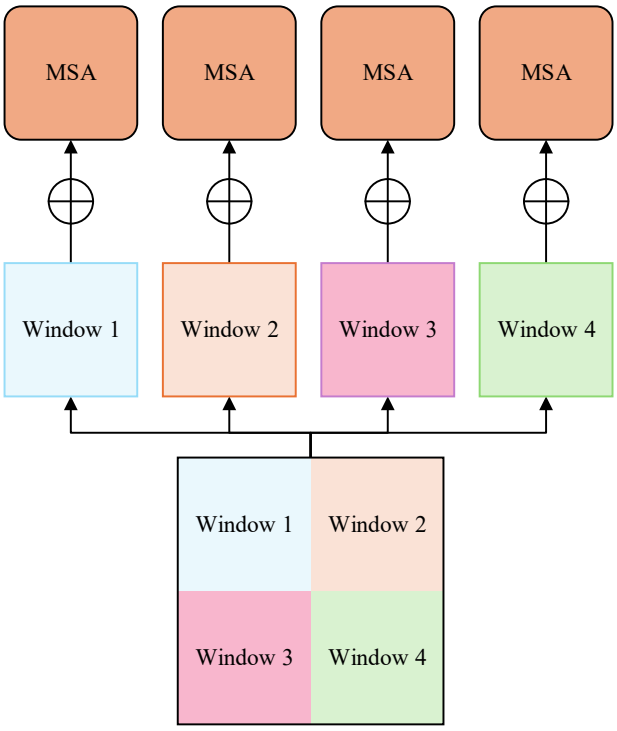
in significant consumption of computing resources; doubling the resolution of an image to be processed will result in resource consumption that is quadruple that of the previous resolution. To address this issue, WMSA can greatly reduce resource consumption. As shown in Figure 3, it divides feature maps into multiple windows and then performs MSA on each window. Given a feature map $X \in R^{C\times H\times W}$ with height H, width W and channel number C. This feature map is divided into non-overlapping windows of height and width M. The size of each window is $M \times M$. For window i, the size of each feature map is $X^i \in R^{M^2\times C}$, and it only needs to perform self-attention mechanism on each window. The computational complexity of MSA and WMSA are shown below, where H, W, C and M represent height, width, channel number, and window size respectively.

$$\Omega(MSA) = 4HWC^2 + 2(HW)^2C \tag{10}$$

$$\Omega(WMSA) = 4HWC^2 + 2M^2HWC \tag{11}$$

From equations (10) and (11), it can be seen that the original MSA model's complexity is a multiple of the square of image height and width, while WMSA has linear multiples with image height and width, where M is a constant. This greatly reduces resource consumption. In this paper, WMSA is added to ViT in order to achieve significant reduction in computational load during operation.

Figure 3 Window-based multi-head self-attention mechanism (see online version for colours)



4.5 Locally enhanced feedforward network

Local information as well as global information for clothing images are both important for colour coordination. Although ViT can effectively capture long-range dependencies, it struggles to utilise local information. However, local information is critical for inferring the features of pixels at specific positions because the model can use surrounding pixels for inference. This ability to extract local information is precisely a strength of CNNs. Therefore combining ViT with CNN allows their advantages to be fully utilised. For this purpose, this paper introduces LeFF by incorporating convolution into the feed-forward network in order to enhance colour coordination capability for clothing images. The computational formula of LeFF is as follows:

$$\begin{cases} 2D\text{-}F = Seq2F\left(X_{Seq}\right) \\ mid\text{-}F' = GeLU(DropPath(1\times1Conv(2D\text{-}F))) \\ mid\text{-}F'' = GeLU\left(DropPath\left(1\times1Conv\left(3\times3Conv\left(mid\text{-}F'\right)\right)\right)\right) \\ 1D\text{-}Seq = Fea2Seq\left(mid\text{-}F''\right) \\ X_{out} = 2D\text{-}F \oplus 1D\text{-}Seq \end{cases} \tag{12}$$

First, the *Seq2F* linear layer is used to transform the sequence X_{seq} into a 2D feature map *2D-F*. Then, 1 × 1 convolution 1 × 1*Conv* is applied to increase the feature dimension, yielding *mid-F'*. Next, 3 × 3 convolution 3 × 3*Conv* is employed to further extract local information from the image. Subsequently, 1 × 1 convolution is used to reduce the feature dimension, obtaining *mid-F''*. The linear layer is then utilised again to adjust *Fea2Seq* so that its dimension matches the input dimension 1*D-Seq*. Finally, the output X_{out} is obtained via a residual connection. Additionally, to enhance the nonlinear relationships between the layers of the neural network, a GeLU activation function is applied after each linear layer and convolutional layer.

4.6 Loss function

The loss function can effectively evaluate the difference between the model's pairing results and actual values; its selection plays an essential role in training the model. Therefore, choosing a suitable loss function is crucial for the performance of the model, as different tasks should adopt different loss functions. In this paper's clothing image colour coordination task, Charbonnier loss (Jayasurya et al., 2025) is used. Compared to L1 loss, its loss curve is smoother. Additionally, due to the inclusion of a small constant near zero, even when gradients become very small, gradient vanishing will not occur. Moreover, when the gradient approaches zero, it does not become too large because of the square root principle, thus avoiding gradient explosion. Its loss function is as follows:

$$\ell\left(I', \hat{I}\right) = \sqrt{\left\|I' - \hat{I}\right\|^2 + \varepsilon^2} \tag{13}$$

where I' is the real clothing image, $\hat{I}$ is the colour-coordinated clothing image after pairing, and $\varepsilon = 10^{-3}$ is a constant.

5 Simulation of fashion colour coordination

5.1 Dataset and experimental parameter settings

The clothing images used in this paper are sourced from the large-scale fashion dataset Deep-Fashion (Jung et al., 2025). This dataset encompasses images from diverse origins, including various shopping platforms and social networks, totaling 280,000 images and 42,369 colour combinations. For experimentation, 22,000 images representing 1,069 colour categories were extracted. These were divided into an 80% training set and a 20% test set. During data processing, each category was sampled to maintain roughly equal image counts. Similar categories were merged, minimising the occurrence of categories with high colour similarity.

The experimental environment configuration for this paper involves both hardware and software aspects. The hardware environment is based on a 64-bit Ubuntu 21.04 system, featuring CUDA version 11.4, cuDNN version 8.4, and Python version 3.8. It is equipped with an NVIDIA RTX-5000 GPU with 16 GB of dedicated graphics memory and 128 GB of host memory. For software configuration, mainstream package management tools such as Conda and Pip can be utilised for installation. To enhance computational efficiency, the neural network framework selected PyTorch 1.0+. It is particularly important to note that the PyTorch version must be compatible with the CUDA version. For model parameter optimisation, AdamW was chosen for parameter updates, with a weight decay of 0.05 and a training cycle of 300 iterations. Each training batch consists of 64 clothing images. The learning rate update employs a cosine annealing strategy, with the minimum learning rate set to 0.01.

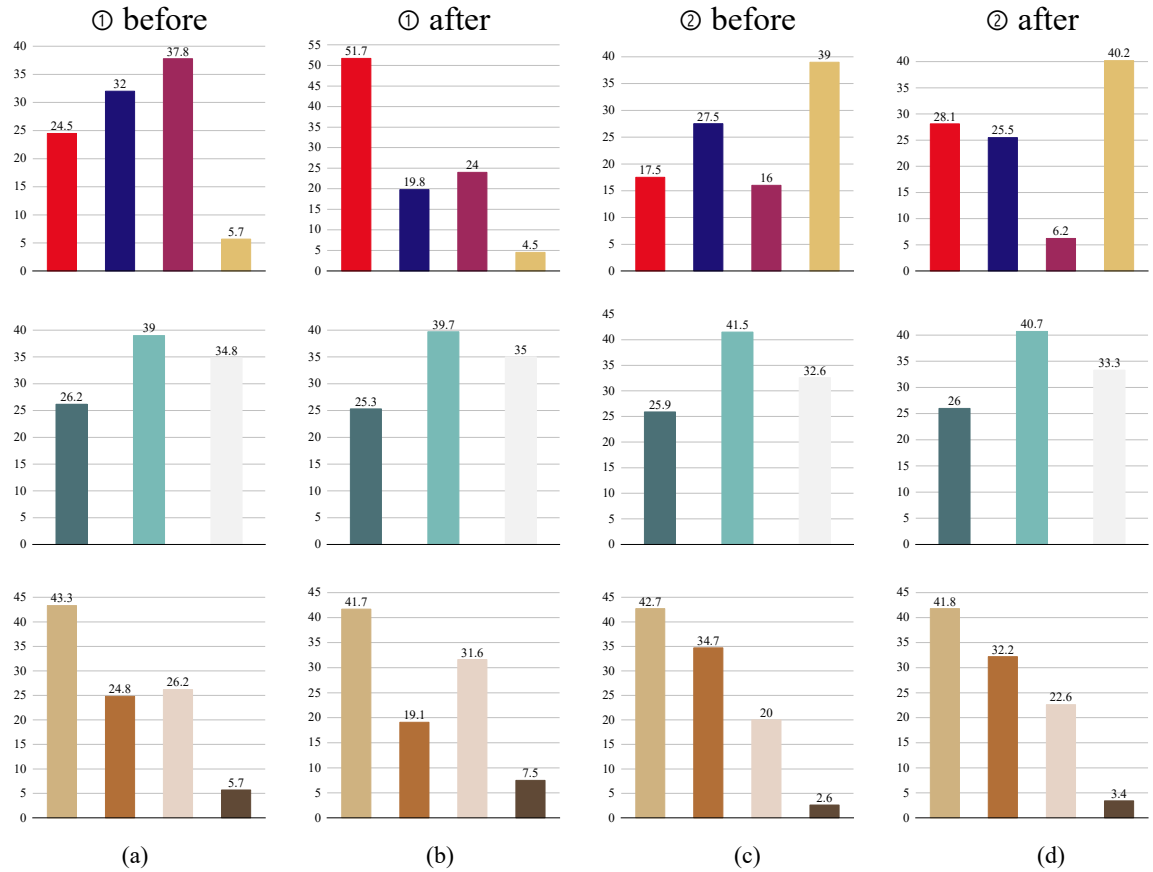
5.2 Analysis of harmonious colour coordination

To further demonstrate the colour harmony effects achieved by UNViT, this section selected two areas within the dataset where pixel colour input significantly influences the surrounding tones as key focus zones for apparel colour coordination. To further illustrate the impact of input pixel colours on the matching results within these zones, this section statistically analysed the colour proportion distribution across various hues within these key areas. The statistical results are presented in Figure 4. The statistical equation is as follows:

$$g_i = \frac{p_i}{p_{sum}} \times 100\% \tag{14}$$

where g_i represents the proportion of the i^{th} colour in the region, p_i denotes the number of pixels occupied by the i^{th} colour in the region, and p_{sum} is the total number of pixels in that region.

Figure 4 The result of harmonious colour coordination (see online version for colours)



Figures 4(a) and 4(c) present the colour statistics results for two locally enlarged regions, while Figures 4(b) and 4(d) show the colour statistics results for the other two locally enlarged regions. By comparing the results in Figures 4(a) and 4(b) with those in Figures 4(c) and 4(d) respectively, we can determine the influence of the input pixel colour on the colour effects within each region. For example, the first row of Figure 4 displays the colour proportion statistics for the two regions. The input pixel colours are two red pixels within these regions. Before inputting pixel colours, the colour proportions in Figure 4(c) column area were: red 24.5%, blue 32%, purple 37.8%, yellow 5.7%. After colour adjustment using the input pixel colours, the colour proportions changed to red 51.7%, blue 19.8%, purple 24%, yellow 4.5%. It can be observed that the proportion of red in the region has increased significantly, while the proportions of blue and purple have decreased substantially. In summary, after inputting the colour of a local pixel point, the colour combination direction generated by UNViT aligns with the colour of the input local pixel point. These results demonstrate that colour combinations can be locally adjusted based on the input pixel colour, highlighting the effectiveness of the UNViT network in apparel colour coordination.

5.3 Comparative experiment

To demonstrate the effectiveness of UNViT, comparative experiments were conducted against benchmark models including DCNN (He et al., 2025), TFGAN (Yan et al., 2023b), GCCAR (Zhang et al., 2023), MPViT (Zhao et al., 2024), and PCSViT (Zou et al., 2025). Multiple objective evaluation metrics were selected for assessment, including peak signal-to-noise ratio (PSNR), SSIM, CF, learned perceptual image similarity (LPIPS), frame-to-frame image distance (FID), and PR curves. The experimental results are shown in Table 1. Higher values of PSNR, SSIM, and CF indicate better clothing colour coordination. Lower values of LPIPS and FID also indicate better clothing colour coordination. On the comprehensive performance metrics PSNR and SSIM for colour coordination, UNViT achieved values of 33.61 and 0.958 respectively, representing improvements of at least 13.62% and 5.04% compared to the other five models. When comparing colour diversity metrics, CF directly reflects the richness of colours in the coloured images. UNViT achieved a CF score of 40.14, surpassing DCNN, TFGAN, GCCAR, MPViT, and PCSViT by 38.18%, 25.91%, 24.81%, 16.28%, and 11.66%, respectively. When compared to LPIPS and FID, UNViT achieved values of 0.16 and 5.09 respectively, representing reductions of at least 15.79% and 21.93% compared to the baseline model. The UNViT model demonstrates high-quality clothing colour coordination capabilities.

The PR curves for different models are shown in Figure 5. The PR accuracies of DCNN, TFGAN, GCCAR, MPViT, PCSViT, and UNViT are 0.748, 0.764, 0.835, 0.851, 0.871, and 0.926, respectively. UNViT achieves a 23.8%, 21.2%, 10.9%, 7.8%, and 6.3% improvement in PR accuracy compared to DCNN, TFGAN, GCCAR, MPViT, and PCSViT, respectively. UNViT employs a dual-jump connection to facilitate feature fusion, ensuring the effective transmission and utilisation of both low-level and high-level features. Secondly, the WMSA designed within ViT significantly reduces computational complexity while capturing long-range dependencies. Furthermore, UNViT incorporates a CNN-based forward network to enhance local information capture, substantially improving the accuracy of clothing colour coordination.

Table 1 Image colour coordination performance metrics results

Model	*PSNR*	*SSIM*	*CF*	*LPIPS*	*FID*
DCNN	17.05	0.802	29.05	0.49	21.89
TFGAN	20.63	0.829	31.88	0.34	19.97
GCCAR	21.79	0.854	32.16	0.26	13.05
MPViT	23.14	0.868	34.52	0.21	8.96
PCSViT	29.58	0.912	35.95	0.19	6.52
UNViT	33.61	0.958	40.14	0.16	5.09

Figure 5 The PR curves for different models (see online version for colours)

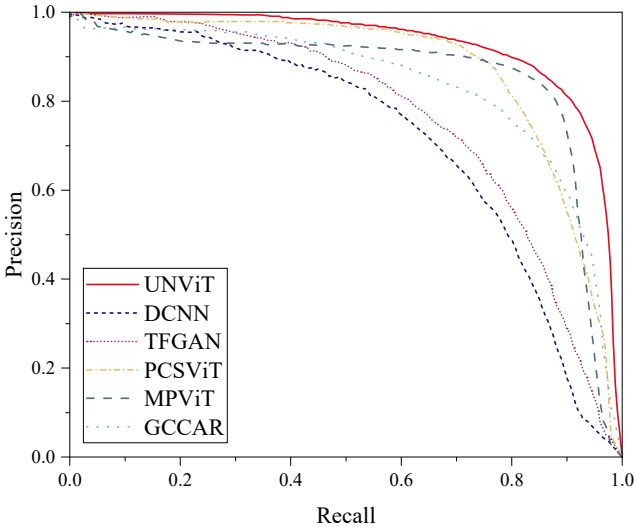


Table 2 Results of the DM statistical test

The suggested model	*Comparison model*	*DM test results*	*p-value*
UNViT	DCNN	29.61	0.00125
	TFGAN	25.47	0.00319
	GCCAR	22.03	0.00752
	MPViT	16.74	0.00894
	PCSViT	12.69	0.00912

Observe the stability of the UNViT model and other models using the DM test values in Table 2. In the DM test, if the p-value is less than 0.05, it is generally considered that there is a significant difference in predictive performance between models. The smaller the p-value, the more significant the difference between models. UNViT exhibits p-values below 0.05 compared to DCNN, TFGAN,

GCCAR, MPViT, and PCSViT, indicating that the pairing results of the UNViT model differ significantly from those of the other five models. The pairing results of the UNViT model are significantly superior to those of the comparison models.

5.4 Ablation experiments for each component in the UNViT model

To validate the effectiveness of each component in the UNViT model, this paper conducts ablation experiments. To assess the capabilities of individual components, the model's colour coordination performance is tested after adding and removing sub-structures from the dataset. Removing ResNet from the UNViT model is denoted as UNViT/Res. Removing the convolutional modules from the ViT component is denoted as UNViT/Cov. Replacing the skip connections with standard concatenation is denoted as UNViT/ST. Replacing the WMSA with MSA is denoted as UNViT/WMSA. The effectiveness of each component is evaluated by comparing PSNR, SSIM, CF, LPIPS, and FID metrics between colour-matched garment images and their original counterparts.

As shown in Table 3, during the colour coordination process, the use of ResNet resulted in a 19.57% reduction in FID, a 0.38 decrease in LPIPS, a 7.93 increase in PSNR, a 0.064 improvement in SSIM, and an 11.53% rise in CF compared to models without ResNet. This demonstrates that ResNet plays a crucial role in clothing colour coordination and significantly enhances the model's colour coordination performance. After incorporating a CNN into the transformer block, LPIPS and FID decreased by 0.15 and 9.3 respectively, while PSNR, SSIM, and CF increased by 13.05, 0.064, and 20.61 respectively. These results demonstrate the CNN's enhancement of colour coordination performance, confirming its role as a crucial component of the model. Comparing the impact of skip connections on the model, incorporating skip connections into UNViT reduced LPIPS and FID by 0.46 and 22.82, respectively, while improving PSNR, SSIM, and CF by 5.6, 0.04, and 8.1, respectively. Comparing metrics across models incorporating WMSA, UNViT achieved lower LPIPS and FID than UNViT/WMSA while outperforming it in PSNR, SSIM, and CF. Overall analysis indicates that the UNViT model with all components integrated delivers the optimal colour coordination performance.

Table 3 Ablation experiment results

Model	*PSNR*	*SSIM*	*CF*	*LPIPS*	*FID*
UNViT/Res	25.68	0.894	28.61	0.54	24.66
UNViT/Cov	20.56	0.817	19.53	0.31	14.39
UNViT/ST	28.01	0.918	32.04	0.62	27.91
UNViT/WMSA	22.36	0.852	21.09	0.39	19.52
UNViT	33.61	0.958	40.14	0.16	5.09

6 Conclusions

Colour coordination is a core factor determining the aesthetic value and market appeal of clothing design. Addressing the current research gap where neglecting local image details leads to suboptimal colour matching quality, this paper proposes a visual transformer-based simulation method for harmonising colour coordination. At the model construction level, this paper explicitly positions ViT within a simulation framework that mirrors human colour perception and designer decision-making processes. By implementing information injection of quantised features through spatial normalised residual networks in the decoder, the method enhances generative diversity while mitigating artefacts. The feature matching mechanism introduced in the discriminator guides the encoder to learn more consistent features, thereby strengthening the model's understanding of apparel colours.

The ViT-based colour matching model achieves breakthroughs in simulation modelling through four core designs. First, dual skip connections facilitate effective fusion between low-level detail features and high-level semantic features, ensuring feature propagation from local to global scales in the design workflow. The windowed multi-head self-attention mechanism reduces computational complexity while accurately capturing long-range colour dependencies. The CNN forward network enhances local information capture capabilities, adapting to detailed processing requirements like garment textures and embellishments. The ResNet module improves model robustness against noise and network stability, preventing performance degradation in complex scenarios.

Experimental results demonstrate that the proposed model exhibits outstanding performance in simulating colour coordination for fashion apparel. With a CF score of 40.14 and an SSIM of 0.958, it significantly outperforms the baseline model, intuitively reflecting the superior harmony and visual quality of its generated colour schemes. Beyond achieving effective colour coordination at the technical level, this model crucially simulates the complex decision-making process of colour pairing in fashion design from a simulation modelling perspective, offering valuable insights for interdisciplinary research at the intersection of fashion design and computer science.

Future research will deepen the simulation modelling of colour coordination in fashion apparel across three dimensions. First, we will explore lightweight ViT architecture design to enhance the model's deployment capability on lightweight devices such as mobile terminals, meeting real-time design demands. Second, we will introduce cross-modal learning mechanisms to integrate multi-dimensional information including garment styles, fabric materials, and wearing scenarios, achieving collaborative optimisation of design elements. Finally, this research will be extended to industrial applications like virtual fitting rooms, strengthening the integration between the model and actual design workflows.

Declarations

The author declares that she has no conflicts of interest.

References

Cao, J., Xu, P., Wu, S., Jiang, W., Lin, R. and Zhang, L. (2025) 'Automatic exploration and transfer design of associative rules in She Ethnic clothing coloration', *Multimedia Tools and Applications*, Vol. 84, No. 12, pp.10269–10289.

Cao, M. (2025) 'Machine learning-based multidimensional sentiment visualisation and analysis of digital media', *International Journal of Information and Communication Technology*, Vol. 26, No. 21, pp.39–54.

Chen, G., Zhang, G., Yang, Z. and Liu, W. (2023) 'Multi-scale patch-GAN with edge detection for image inpainting', *Applied Intelligence*, Vol. 53, No. 4, pp.3917–3932.

Dong, Z-J., Liang, J-F., Zhang, Z-J. and Wei, S-S. (2023) 'The perceptual evaluation of clothing sustainable color in clothing design', *Journal of Fiber and Bioengineering and Informatics*, Vol. 16, No. 3, pp.229–241.

Gu, X., Huang, J., Wong, Y., Yu, J., Fan, J., Peng, P. and Kankanhalli, M.S. (2023) 'PAINT: photo-realistic fashion design synthesis', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 2, pp.1–23.

Guerra, F.d.C.F. and Mota, W.S. (2006) 'Current transformer model', *IEEE Transactions on Power Delivery*, Vol. 22, No. 1, pp.187–194.

He, Z., Tan, Y. and Li, S. (2025) 'A new CNN deep learning model for computer-intelligent color matching', *Nonlinear Engineering*, Vol. 14, No. 1, pp.20–38.

Hu, Z-H., Li, X., Wei, C. and Zhou, H-L. (2019) 'Examining collaborative filtering algorithms for clothing recommendation in e-commerce', *Textile Research Journal*, Vol. 89, No. 14, pp.2821–2835.

Huang, T. (2020) 'Contrast and color combination of tint in aesthetic appreciation of clothes', *Revista Argentina de Clínica Psicológica*, Vol. 29, No. 2, pp.45–57.

Jayasurya, S., Geetha, S., Abdullah, A.S. and Mishra, U. (2025) 'UWE-Net: a deep learning framework for underwater image enhancement integrating CBAM and Charbonnier loss', *Procedia Computer Science*, Vol. 258, pp.689–698.

Jung, J., Kim, H. and Park, J. (2025) 'Deep fashion designer: generative adversarial networks for fashion item generation based on many-to-one image translation', *Electronics*, Vol. 14, No. 2, pp.22–36.

Li, H-C., Wang, L-K., Chang, Y-K. and Huang, K-Y. (2025) 'Establishing colour harmony evaluation and recommendation model for clothing colour matching based on machine learning and deep learning', *Fashion and Textiles*, Vol. 12, No. 1, pp.27–41.

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z. and Loy, C.C. (2024) 'Transformer-based visual segmentation: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, pp.10138–10163.

Mrinali, A. and Gupta, P. (2025) 'Cloth-Net: improved hybrid adversarial network with dense vision transformer for 2D-3D image classification for accurate cloth recommendation engine', *International Journal of Clothing Science and Technology*, Vol. 37, No. 3, pp.402–442.

Park, T., Liu, M-Y., Wang, T-C. and Zhu, J-Y. (2019) 'Semantic image synthesis with spatially-adaptive normalization', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2337–2346.

Sang, H. (2025) 'Mask-embedded transformer for English text recognition and correction', *International Journal of Information and Communication Technology*, Vol. 26, No. 35, pp.1–17.

Shamoi, P., Inoue, A. and Kawanaka, H. (2020) 'Modeling aesthetic preferences: color coordination and fuzzy sets', *Fuzzy Sets and Systems*, Vol. 395, pp.217–234.

Williams, C., Falck, F., Deligiannidis, G., Holmes, C.C., Doucet, A. and Syed, S. (2023) 'A unified framework for U-Net design and analysis', *Advances in Neural Information Processing Systems*, Vol. 36, pp.27745–27782.

Xu, M. (2025) 'Image colorization based on transformer', *Scientific Reports*, Vol. 15, No. 1, pp.31–49.

Xu, W., Fu, Y-L. and Zhu, D. (2023) 'ResNet and its application to medical image processing: research progress and challenges', *Computer Methods and Programs in Biomedicine*, Vol. 240, pp.97–110.

Yan, H., Zhang, H. and Zhang, Z. (2023a) 'Learning to disentangle the colors, textures, and shapes of fashion items: a unified framework', *IEEE Transactions on Multimedia*, Vol. 26, pp.5615–5629.

Yan, H., Zhang, H., Shi, J., Ma, J. and Xu, X. (2023b) 'Toward intelligent fashion design: a texture and shape disentangled generative adversarial network', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 19, No. 3, pp.1–23.

Yum, M. (2023) 'Digital image color analysis method to extract fashion color semantics from artworks', *Multimedia Tools and Applications*, Vol. 82, No. 11, pp.17115–17133.

Zhang, Q. and Yang, Y-B. (2021) 'Rest: an efficient transformer for visual recognition', *Advances in Neural Information Processing Systems*, Vol. 34, pp.15475–15485.

Zhang, Z., Sun, L., Yang, Z., Chen, L. and Yang, Y. (2023) 'Global-correlated 3D-decoupling transformer for clothed avatar reconstruction', *Advances in neural information processing systems*, Vol. 36, pp.7818–7830.

Zhao, T., Li, G. and Zhao, S. (2024) 'End-to-end image colorization with multiscale pyramid transformer', *IEEE Transactions on Multimedia*, Vol. 26, pp.11332–11344.

Zhao, X., Yang, H., Shi, X., Liu, K., Wang, Y. and Zhang, G. (2023) 'CPRM: color perception and representation model for fabric image based on color sensitivity of human visual system', *Textile Research Journal*, Vol. 93, No. 13, pp.2956–2970.

Zou, X., Peng, Y., Li, G. and Cao, X. (2025) 'PCSViT: efficient and hardware friendly pyramid vision transformer with channel and spatial self-attentions', *Neurocomputing*, Vol. 636, pp.12–27.