Design Prior Guided Semantic Component Transformer for Hand-Drawn Fashion Sketch Rendering

Ning Wang[1], Anqi Zou[1], Shuge Qian[1], Baoli Sun[1], Zhiyong Wang[2], Zhihui Wang[1*]

[1*]Software of Technology, Dalian University of Technology, Tuqiang Street, Dalian, 116000, LiaoNing, China.
[2]The School of Computer Science, The University of Sydney, Darlington Campus, City Rd, Sydney, 2006, New South Wales, Australia.

*Corresponding author(s). E-mail(s): zhwang@dlut.edu.com;
Contributing authors: nwang@dlut.edu.com; bigben@mail.dlut.edu.cn; qsugar@mail.dlut.edu.cn; baoli@mail.dlut.edu.cn; zhiyong.wang@sydney.edu.au;

Abstract

Fashion sketch rendering is a fundamental process that efficiently transforms a conceptual fashion sketch into an illustration conforming to design aesthetics. Recently, many deep learning-based methods have advanced sketch rendering through pixel-aware or region-aware perception. However, they often struggle with obtaining accurate semantic-region-aware perception of hand-drawn fashion sketches, leading to difficulties in capturing design priors like fashion overall color aesthetics and clothing component independence, resulting in dull rendering and fabric pattern bleeding problems in rendering output. To address these challenges, we propose a Design Prior Guided Semantic Component Transformer (DPG-SCT) to take sketch semantics into account for capturing design priors. This marks the first in-depth exploration into the rendering of hand-drawn fashion sketches. The semantic-region-aware perception, achieved through several Semantic Component Transformer (SCT) blocks, leverages semantic component regions derived from a large-scale segmentation model. To be specific, SCT effectively captures design priors by comprehending the distinct semantic regions and their contextual relationships within a sketch, which is achieved through component-wise linear projection and feature propagation. We further devise two losses to enhance understanding of fashion design priors: the Fashion Color Prior

loss, which fosters a multimodal color distribution for overall compliance of color aesthetics, and the Fashion Semantic Component Prior loss which ensures the independence and fidelity of each clothing component. Comprehensive experimental results on the SketchCouture dataset demonstrate that our proposed method can produce high-quality sketch rendering results that comply with the technical demands and creativity expectations of professional designers.

Keywords: Design Prior, Semantic-region, Sketch Rendering, Transformer

1 Introduction

Fashion sketch rendering aims to automatically generate creative design illustrations for a given sketch or a sparse hand-drawn fabric swatch. It has a wide range of applications in the fields of artistic design, and garment production, serving as a source of inspiration for designers. Traditional hand-drawn renderings, while artistically valuable, are labor-intensive and limited in quantity, often falling short in industrial production demands. Consequently, to enhance efficiency and better fulfill the evolving aesthetic demands, there is a growing exploration into deep learning-based hand-drawn sketch rendering [1, 2].

Recent developments on reference-based hand-drawn sketch rendering have extensively utilized generative models [3]. TextureGAN [4], utilizes a two-stage training approach to achieve patch-based perception in image synthesis. Yan *et al.* [5] developed a Spatially Corresponding Feature Transfer (SFCT) module to establish dense semantic correspondences at pixel level for reference. Li *et al.* [6] mitigated gradient conflict issues of SFCT by exploring a stop-gradient attention mechanism. Furthermore, DPGAN [7] emphasized the significance of incorporating drawing prior within a dual-color space. In a distinct approach, the PITI model [8] adapted a pre-trained large diffusion model to facilitate image-to-image translations in a pixel-aware manner. All these methods are implemented in a rendering manner either through pixel-aware perception [5–8] or through region-aware perception [4].

Despite the notable advancements achieved in the field of sketch rendering[9, 10], the specific area of fashion sketch rendering remains underexplored. A major limitation in existing methods is their lack of semantic-region-aware perception, which hinders the effective understanding of fashion design priors, such as color matching and the distinctiveness of hand-drawn fabric patterns within the components.

Therefore, to address the challenge, we introduce the Design Prior Guided Semantic Component Transformer (DPG-SCT) tailored for fashion sketch rendering. Technically, our approach begins by using a large-scale segmentation model to obtain semantic components in fashion sketches accurately, such as bag, and skirt suit semantic components. To enhance the sensitivity of the input features to semantic component regions, our DPG-SCT architecture comprises an encoder and a decoder, each constructed with their respective Semantic Component Transformer (SCT) blocks. The encoder's SCT blocks facilitate semantic-region-aware component region perception through component-wise linear projection and feature propagation to obtain a

semantic-region-aware component feature map. The decoder's SCT utilizes two modulation factors to optimize image rendering quality: a high-level latent code from the encoder for global uniform coordination and the semantic-region-aware component feature map for modulating local fabric pattern details. Based on these, DPG-SCT ensures both intra-component independence and inter-component harmony in the illustrations. We also introduce two innovative loss functions: Fashion Design Prior(FDP) loss is introduced to regulate color distribution in the generated images to follow a multimodal distribution, preventing excessively uniform color distribution that may lead to overly complex color combinations, and Fashion Semantic Component Prior (FSCP) loss to maintain consistent hand-drawn fabric patterns within components.

In summary, the key contributions of this work are as follows:

- We propose DPG-SCT, the first fashion sketch rendering by utilizing a large-scale segmentation model for precisely acquiring semantic component regions within sketches.
- We devise an SCT block to leverage these semantic component regions effectively, which facilitates semantic-region-aware perception through local and global modulations, ensuring that each component is distinct yet harmoniously integrated into the overall aesthetics.
- We develop two prior guided losses: the FDP Loss to constrain the generated color distribution adherence to a multimodal distribution, preventing excessive color diversity, and the FSCP Loss to maintain pattern consistency within individual clothing components.
- We constructed a new dataset SketchCouture comprising hand-drawn sketch-fashion design illustrations. Our extensive experimental results show that our proposed DPG-SCT excels in creating fashion illustrations that not only are visually appealing but also demonstrate a harmonious color balance with creative distinction in fashion components.

2 Related work

In the past few years, numerous works have been proposed regarding sketches [11, 12], meanwhile, many image rendering techniques have been developed for various sketches, including fashion sketch [4, 5, 13, 14], anime sketch [7, 15–19], icon sketch [20–22] and scenery sketch [23–25].

Fashion Sketch: Existing fashion sketch rendering methods aim to generate realistic images of individual fashion components. Cui *et al.* [13] proposed FashionGAN to synthesize an image with specified fabric patterns by designing a two-to-one GAN architecture. Xian *et al.* [4] allowed users to place a texture patch on a sketch at arbitrary locations and generated images with assigned texture. Yan *et al.* [5] introduced a novel framework for synthesizing individual clothing components, encompassing both sketch generation and rendering processes. The sketch-generation network produces a variety of component sketches by using a GAN-based module, and a multi-conditional feature interaction module to generate textures. Wu et al. [14] developed the StyleMe

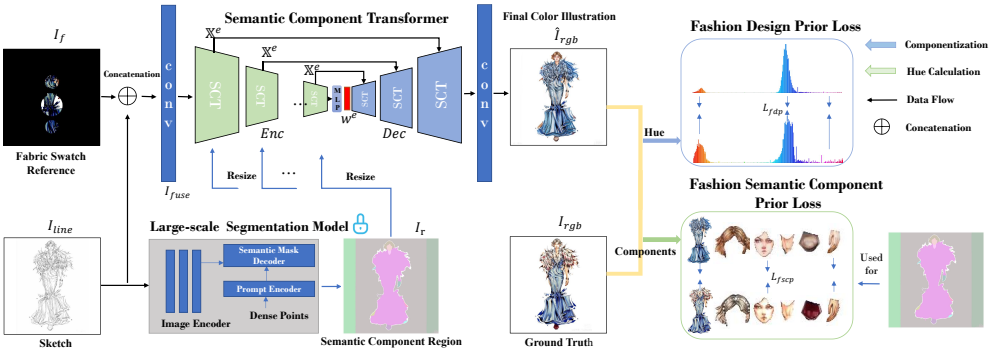

Fig. 1 Illustration of the proposed DPG-SCT method. First, a fashion sketch I_{line} and a fabric swatch reference I_f are first concatenated and fed into a conv layer to form I_{fuse}. Second, I_{fuse} goes through several Semantic Component Transformer (SCT) blocks, concluding with a conv layer that generates the color illustration $\hat{I}_{rgb}$. Note that I_{line} is also processed by a large-scale segmentation model to extract semantic component regions I_r, which help Enc's SCT learn semantic-region-aware feature maps $\mathbb{X}^e$. Then, we apply MLP to get the global latent code w^e. Two fashion prior losses L_{fdp} and L_{fscp} help produce $\hat{I}_{rgb}$ towards multimodal distribution and ensure independent hand-drawn fabric pattern in each component.

system, which synthesizes final design illustrations from designers' sketches and reference images via a GAN equipped with channel attention. While these methodologies may excel in generating individual components, they do not extensively explore the synthesis and integration of various components within full-body, hand-drawn fashion designs, which is the focus on our proposed method.

Anime: Furusawa *et al.* [15] and Sangkloy *et al.* [26] developed a color-guided rendering model for manga images. Ci *et al.* [16] proposed a pre-trained local feature network to extract the semantic features of a line art, thereby eliminating the problem of over-fitting in a certain type of line arts and generating accurate shading. Considering the prior for anime image, Dou *et al.* [7] proposed to generate more attractive images with harmonious color composition and fewer artifacts by exploring the drawing prior in HSV color space. Zhang *et al.* [17] proposed a split filling mechanism for flat anime by explicitly controlling the 'influence areas' of the user color hints. However, these methods generally focus on rendering quality without exploring design priors that are specific to the fashion design industry.

Icon Sketch: Icon-Gan [20] colorized icon sketches by introducing a dual conditional generative adversarial network. Li *et al.* [21] presented a flat color rendering network for icons, by featuring a style-structured disentangled module and normalizing flow for enhanced icon quality. Due to the simplicity of icon sketches, these methods do not apply to fashion sketches.

Scenery Sketch: PITI [8] adapted a pre-trained large diffusion model to facilitate image-to-image translations. ControlNet [9], StyleAdapter [27], FastComposer [28], UniCanvas [29] and UniControl [10] utilized the capabilities of large-scale models for text-to-image generation, producing high-quality colorization with low network resource usage. However, these diffusion-based models rely on textual references, leading to uniform color in fashion images generated from default or absent prompts.

In summary, the methods mentioned above are not applicable to produce images that exhibit overall fashion color aesthetics and component independence needed for full-body, hand-drawn fashion sketch rendering. In addition, they mainly concentrate on refining network architectures or pipelines for pixel-aware [5–8, 13, 16, 30] or region-aware [4, 17, 31, 32] perception, thereby ignore the critical aspect of semantic-region-aware perception in sketch, which is essential for a comprehensive understanding of design priors. Our proposed DPG-SCT method aims to address such limitations with design priors and the semantic component transformer.

3 Proposed method

The overall framework of DPG-SCT is shown in Figure 1. Given the fashion sketch I_{line} and the fabric swatch reference I_f as inputs, I_{fuse} is derived through a Concatenation and Convolution operation. *Enc* utilizes it and $I_r \in \mathbb{R}^{1\times W\times H}$ to generate semantic-region-aware feature maps $\mathbb{X}^e = \{X_1^e, \ldots, X_M^e\}$, where $X_m^e \in \mathbb{R}_m^{c_m^e \times W_m^e \times H_m^e}$ represents the output feature map of the m-th SCT block and M is the number of SCT blocks within *Enc*. Then, we use an MLP to get the latent code w^e(Section 3.1). The decoder *Dec*, following modulation operation in its SCT blocks, outputs a series of decoding feature maps represented as $\mathbb{X}^d = \{X_1^d, \ldots, X_N^d\}$, where $X_n^d \in \mathbb{R}^{C_n^d \times W_n^d \times H_n^d}$ denotes the output of the n-th *Dec* SCT block and N is the number of SCT blocks within *Dec* (Section 3.2). The number of SCT blocks in both *Enc* and *Dec* is equally, denoted by $M = N$. Such symmetry facilitates *Dec* to utilize the semantic-region-aware feature of *Enc*.

3.1 Encoder Phase

The Semantic Component Transformer (SCT) block within *Enc* serves as the fundamental unit of our network. We first use a frozen large-scale segmentation model SAM [34] which is capable of performing accurate segmentation based on sketch information to predict semantic component regions I_r (e.g., skirt region, face region, background region) from I_{line} and utilizing the results of the average scatter sampling method as prompts. I_r is further broadcasted to all SCT blocks by employing scaling factors with nearest neighbor interpolation.

3.1.1 Component Linear Projection

The purpose of Component Linear Projection (CLP) is to produce component embeddings $\mathbb{C}^e \in \mathbb{R}^{c_m^e \times 1}$ based on the semantic component region I_r. The implementation of CLP is as follows by adding up the values of all positions corresponding to I_r^t in $\mathbb{X}^e$:

$$\mathbb{C}_t^e = \frac{1}{N_{I_r^t}} \cdot \sum_{(p,q)\in I_r^t} \mathbb{X}^e(p,q), t \in [1 \ldots T], \tag{1}$$

where I_r^t denotes the t-th region of I_r and T denotes the total number of regions. $(p, q) \in I_t^r$ represents the coordinates of a pixel in I_r^t. $N_{I_t^r}$ indicates the total number of pixels in I_r^t.

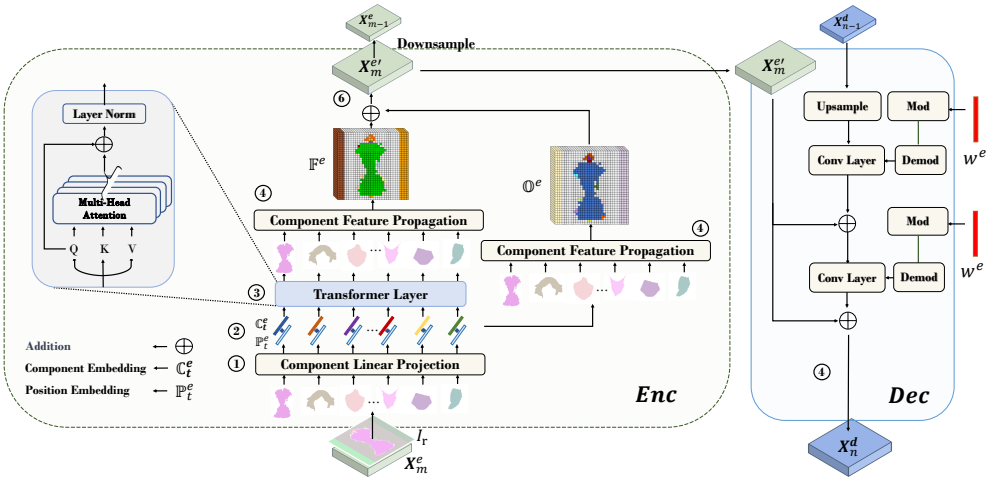

Fig. 2 Illustration of the Semantic Component Transformer block in *Enc* and *Dec*. The process involves: ① using Component Linear Projection (CLP) to extract the component embeddings $\mathbb{C}^e$ from I_r and X_m^e; ② adding to position embeddings $\mathbb{P}^e$ to acquire position-wise component embeddings $\mathbb{CP}^e$; ③ applying Multi-Head Attention mechanism [33] of the Transformer; ④ passing through Component Feature Propagation (CFP) to generate inter-component feature map $\mathbb{F}^e$; ⑤ a parallel branch with CFP, removing the Transformer, to get intra-component feature map $\mathbb{O}^e$; ⑥ adding the outputs of the two branches as the final output $X_m^{e'}$; and ⑦ performing modulation in the SCT block of *Dec* to obtain X_n^d.

Positional encoding allows the model to better capture relative positional information among components. Hence, we employ the sinusoidal positional encoding function [33] to acquire position-wise component embedding $\mathbb{P}^e \in \mathbb{R}^{c_m^e \times 1}$. The final output can be expressed as:

$$\mathbb{CP}_t^e = \mathbb{C}_t^e + \mathbb{P}_t^e, t \in [1...T]. \tag{2}$$

3.1.2 Multi-Head Attention Mechanism

Due to the need to guarantee the fashion's overall color aesthetics, we employ a multi-head attention mechanism to capture the correlations among these components. The position-wise component embeddings $\mathbb{CP}^e \in \mathbb{R}^{c_m^e \times T}$ is encoded into query matrix Q, key matrix K, and value matrix V:

$$Q = \mathbb{CP}^e * W^Q, K = \mathbb{CP}^e * W^K, V = \mathbb{CP}^e * W^V, \tag{3}$$

where W^Q, W^K, W^V are learnable matrices.

The multi-head attention formula is represented as follows:

$$\text{Head}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \tag{4}$$

where $Head \in \mathbb{R}^{c_m^e \times T}$ encompassing feature correlations between components. $Head = (head_1, head_2, ..., head_h)$, where h denotes the number of attention heads. By

concatenating the individual heads together and multiplying them with the learnable matrix W^o, we obtain the output of the multi-head attention:

$$M{=}\text{concat}(\,head_1,\ldots,head_h)W^o, M{\in}\mathbb{R}^{c_m^e \times T} \tag{5}$$

The model leverages a multi-head attention mechanism to concurrently learn varied representations across multiple subspaces, enhancing its ability to discern diverse features and relationships among components. Each attention head concentrates on distinct aspects, enabling a comprehensive characterization of inter-component context, which is crucial for coherent integration across the component regions.

3.1.3 Component Feature Propagation

After obtaining the multi-head attention feature between each component region M, we proceed with the component feature propagation (CFP) to ensure consistency in feature representation within the same component region. The inter-component feature map implementation is as follows:

$$\mathbb{F}_t^e(p,q){=}MH_t, \quad (p,q){\in}I_t^r, \quad t{\in}[0...T], \tag{6}$$

which represents M_t are propagated to their corresponding local component region I_r^t.

To enhance the robustness of the semantic-region-aware feature map, it's essential to maintain the original component feature embedding. Hence, we also perform CFP to generate the original component feature map $\mathbb{O}^e$ by adding another branch but removing the Transformer layer. The implementation is as follows:

$$\mathbb{O}_t^e(p,q){=}\mathbb{CP}_t^e, \quad (p,q){\in}I_r^t, \quad t{\in}[0...T]. \tag{7}$$

These two features map $\mathbb{O}^e$ and $\mathbb{F}^e$ are then fused as the final semantic-region-aware component feature map. Finally, we apply a multilayer perception P to get the global latent code $w^e \in \mathbb{N}(0, I)$:

$$X_m^e{=}\mathbb{O}_t^e{+}\mathbb{F}_t^e, \quad t{\in}[0...T]; \quad w^e{=}P(X_m^e) \tag{8}$$

3.2 Decoder Phase

As illustrated in Figure 2, the Semantic Component Transformer (SCT) block within *Dec* receives three inputs: the component region feature map X_m^e derived from the corresponding block in *Enc*; the decoding feature map X_{n-1}^d from the previous SCT block and the global latent code $\boldsymbol{w^e}$.

In pursuit of optimizing image rendering quality and fashion overall aesthetics, the SCT block in *Dec* contains two modulation factors. Inspired by [35], the model controls the impact of $\boldsymbol{w^e}$ on global style and X_m^e to modulate local style details. The operation process unfolds as follows:

$$X_n^d{=}X_m^e{+}\left(\underset{\boldsymbol{w^e}}{\text{Conv}}\left(X_m^e{+}\underset{\boldsymbol{w^e}}{\text{Conv}}\left(\text{Upsample}\left(X_{n-1}^d\right)\right)\right)\right). \tag{9}$$

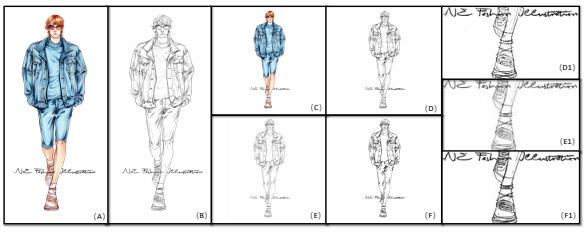

Fig. 3 The examples of sketch-fashion design illustration pairs. **(A)** Super-resolution Illustration [36], **(B)** Sketch extracted from super-resolution illustration [37], **(C)(E)** Illustration and Sketch resized to 512×512 with pure white padding, **(D)** Sketch extracted from the original size illustration, **(F)** Binary Sketch, **(D1)(E1)(F1)** Local enlarged images of (D)(E)(F).

where Conv$_{\boldsymbol{w^e}}$ represents the convolutional operation after weight modulation using $\boldsymbol{w^e}$. X^d_{n-1} denotes the decoding feature map of the previous SCT block. The *dec* ultimately outputs a color illustration $\hat{I}_{rgb}$.

3.3 Loss Function

Fashion Design Prior Loss. One of the fundamental principles of fashion design is to ensure that the color combinations of the clothing components are not overly complicated. To achieve this goal, we introduce a Fashion Design Prior (FDP) Loss, aimed at constraining the color distribution in generated fashion illustrations to a multimodal distribution, countering the complexity implied by a uniform distribution [7] indicating excessive color variety. We use the Hue in the HSV color space, as it describes the color property. The loss function $\mathcal{L}_{fdp}$ is defined as follows:

$$\mathcal{L}_{fdp}=\frac{1}{N}\sum_{i=1}^{N}\left(C_g^h(i)-Y_g^h(i)\right)^2, \tag{10}$$

where $C_g^h(i)$ represents the i-th element of the Hue channel in the generated image, $Y_g^h(i)$ is the center color of the cluster to which the corresponding pixel in the real fashion illustrations belongs, and N denotes the product of image width and height. Besides, we employ the K-means clustering method to acquire the clustering centers for the Hue values of pixels in the image.

Fashion Semantic Component Prior Loss. To guarantee single-type hand-drawn fabric patterns in each component region of generated images, we apply a componentization operation. This process produces a component-wise generated image, denoted as $\hat{I}^t_{rgb}$, and a component-wise real image, represented by I^t_{rgb}.

$$\hat{I}^t_{rgb}=\hat{I}^t_{rgb}(p,q)\in I^t_r, t\in[1...T]; I^t_{rgb}=I^t_{rgb}(p,q)\in I^t_r, t\in[1...T], \tag{11}$$

Next, we align the style of $\hat{I}^t_{rgb}$ and I^t_{rgb} to achieve local style consistency in $\hat{I}^t_{rgb}$, and represent the style information using Gram matrices:

$$G^l_{ij}=\sum_{k=1}^{W\times H}\{F^l_{ik}F^l_{jk}\}, i,j\in[1...C], \tag{12}$$

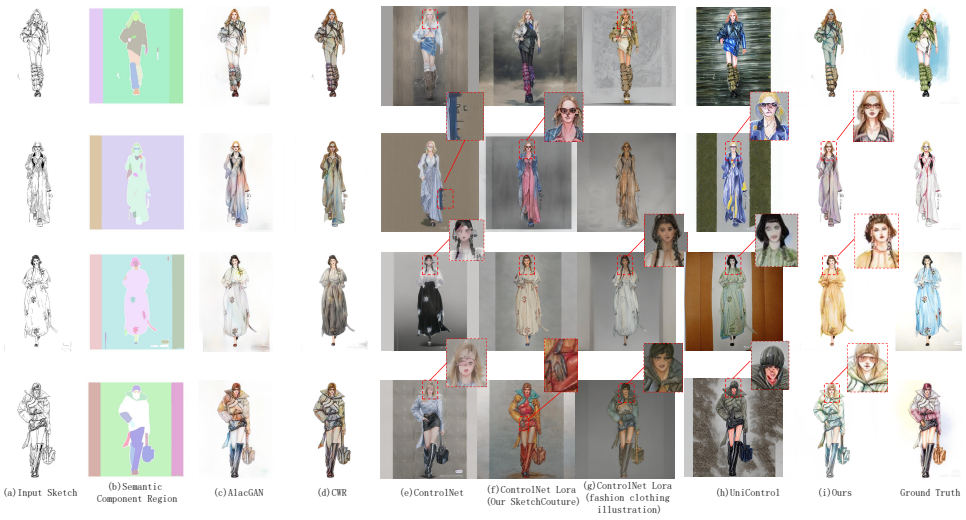

Fig. 4 Visual comparisons with state-of-the-art automatic sketch rendering methods. (a) Input Sketch, (b) Semantic Component Region, (c) AlacGAN [16], (d) CWR [31], (e) ControlNet [9], (f) LoRA [38] with our SketchCouture, (g) the public LoRA with fashion clothing illustration, (h) UniControl [10], (i)Ours. Please zoom in to view the details.

where F_{ik}^{l} and F_{jk}^{l} denote the k-th feature value at the i-th channel and j-th channel, respectively, and C denotes the total number of channels in the feature map. This matrix calculates hidden connections between features to capture statistical information of image styles. Finally, the Fashion Design Component (FSCP) Loss is formulated as:

$$\mathcal{L}_{fscp}=\frac{1}{4N_l^2M_l^2}\sum_{i,j}\left\{\left(G_{ij}^{l}\left(\hat{I}_{rgb}^{t}\right)-G_{ij}^{l}\left(I_{rgb}^{t}\right)\right)^2|t\in[1...T]\right\}, \tag{13}$$

where N represents the number of feature maps, M represents the size of each feature map. Moreover, we adopt an Adversarial loss $\mathcal{L}_a$ to enhance the similarity of the generated images to their real image.

$$\min_{CT}\max_{\mathcal{D}}\mathcal{L}_a=\log(1+\exp(\mathcal{D}(Y_g)))+\log(1+\exp(-\mathcal{D}(C_g))), \tag{14}$$

where $\mathcal{D}$ is a discriminator based on a convolution network, $\hat{I}_{rgb}$ represents the generated fashion image, and I_{rgb} represents the real fashion image.

We also introduce $\mathcal{L}_1$ loss to minimize the absolute differences :

$$\mathcal{L}_1=\|I_{rgb}-\hat{I}_{rgb}\|_1, \tag{15}$$

$\mathcal{L}_1$ loss measures the pixel-level discrepancy between two RGB images, which can easily result in a blurring effect. Our network is optimized by minimizing the following total loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total}=\mathcal{L}_a+\mathcal{L}_1+\mathcal{L}_{fdp}+\mathcal{L}_{fscp}. \tag{16}$$

Fig. 5 Visual comparisons with state-of-the-art reference-based sketch rendering methods. (a) Input sketch, (b) Semantic component region, (c) Fabric reference, (d) SCFT [1], (e) SGA [6], (f) DPGAN [7], (g) TextureGAN [4], (h) PITI [8], (i) Ours, (j) Ground Truth. Please zoom in to view the details.

Table 1 Comparisons with existing reference-based methods.

Method	SCFT[1]	SGA[6]	DPGAN[7]	TextureGAN[4]	PITI[8]	Ours
SSIM↑	0.86	0.85	0.87	0.86	**0.91**	0.90
PSNR↑	19.08	19.70	19.65	22.13	16.27	**22.49**
FID↓	61.38	63.63	46.37	48.84	50.40	**34.83**
LPIPS↓	0.15	0.14	0.13	0.13	0.18	**0.12**
MOS↑	2.10	2.05	2.95	3.40	3.72	**3.92**

4 Experiments

4.1 Datasets

Training Datasets. We constructed and open-sourced a new dataset SketchCouture comprising 11,174 hand-drawn sketch-fashion design illustration pairs with various styles (e.g., marker and watercolor) sourced from the Internet. We have taken measures to respect the copyright of the authors by retaining any signature watermarks on the images, thereby acknowledging their ownership and contribution. As illustrated in Figure 3, to obtain sketches from fashion design illustrations, we first apply a super-resolution technique [36] to the original illustrations (B) and extract the sketches by [37]. As shown in the last column, compared to (D1), our sketches (E1) and (F1) are cleaner and have less noise, demonstrating the importance of super-resolution. The resulting sketches are then resized to 512×512 dimensions with pure white padding (E) and subjected to binarization (F). During the training phase, we randomly selected (E) and (F) as input sketches.

Test Datasets. We evaluated our DPG-SCT on 1,590 hand-drawn sketch-fashion design illustrations: the first, IDA, encompasses 734 authentic real hand-drawn

Fig. 6 Visual ablation comparisons for modules. Please zoom in to view the details.

sketches from the internet for unconditional sketch rendering. The second, IDB, comprises 856 sketches derived from color fashion illustrations using the method described in [37] that do not overlap with the training dataset samples. The procedure for obtaining the fabric swatch reference involves creating a random, irregular mask composed of 1-10 randomly generated shapes, including lines, circles, ellipses, and squares, as outlined in [39]. Subsequently, this mask is multiplied with a reference image to derive the fabric swatch reference.

4.2 Experimental Settings

Our model is implemented in PyTorch and all the experiments are performed on a computer with two Tesla V100 GPUs. The initial learning rate of the Adam optimizer is 0.002 and by multiplying the learning rate by 0.8 after each epoch, the total epoch is 20. The momentum parameters are set as follows: $\beta_1 = 0$, $\beta_2 = 0.99$. The batch size is fixed at 8 throughout the whole training process. All image size is equal to 512×512. The binary threshold for getting sketches is 128. The number of attention heads $h = 8$. The number of segmented regions differs across images, with an average of 14. Our DPG-SCT model accommodates any number of regions. The process for generating SAM regions is as follows: Initially, the semantic regions segmented by SAM are sorted by area. Each distinct object within the segmentation is denoted by an incrementing integer and then populated into a Numpy array. Typically, index 0 signifies the largest segmented semantic region, often corresponding to the background. Leveraging SAM's robust semantic understanding capabilities enables the acquisition of reasonable segmentation component regions. In special trivial and unwanted segmentation cases, we can adapt the point prompts to improve the semantic region segmentation accuracy in sketches, thereby enhancing the rendering results during the testing phase.

Table 2 Comparisons with existing automatic-based methods.

Method	AlacGAN[16]	CWR[31]	ControlNet[9]	UniControl[10]	Ours
FID↓	82.67	64.48	50.82	59.9	**43.23**
MOS↑	2.50	3.15	3.36	3.50	**3.62**

4.3 Metrics

For the test dataset IDB, we use Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity(LPIPS), Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Opinion Score (MOS) as evaluation metrics. Since the test dataset IDA is unpaired, we only compute FID and MOS. The user study we used is MOS, which directly reflects the human perception of the visual quality of our results. First, we randomly select 1000 images from the test set and engage 20 volunteers to rate the results using integer scores ranging from 1 (worst) to 5 (best). Volunteers should consider the reasonableness of color aesthetics and the presence of noticeable artifacts within components, such as pattern bleeding or semantic inconsistencies. Finally, we calculate the mean of these scores to determine MOS for the method.

4.4 Evaluation

4.4.1 Qualitative Evaluation

Reference-based sketch rendering method comparison. In Figure 5, we present qualitative visual comparisons between our proposed method and existing state-of-the-art reference-based models. The use of dense pixel-aware correspondence mapping in SCFT [1] for learning local color references demonstrates a significant influence on the accuracy of the rendering. This is particularly evident when the local fabric swatch references are used. As indicated by the red box in column (d), the fabric swatch reference is a black fabric pattern, while the vast majority of the generated clothing is red. Although the SGA model [6] employs a stop-gradient operation to mitigate the issue of color misalignment, this approach is only partially successful. As evidenced in column (e), highlighted by the red box, similar challenges persist in accurately aligning colors. DPGAN [7] and TextureGAN [4] achieve accurate color alignment, but their lack of semantic-region-aware perception leads to pattern bleeding issues, as shown in the red boxes in columns (f) and (g). Diffusion-based PITI model [8] is capable of generating highly saturated and high-contrast colors. However, it suffers from distortion and geometric deformation, as exemplified by the facial representation within the red magnified box in column (h). Our method is able to capture semantic-region-aware region perception, accomplishing reasonable color combinations in the overall design (e.g., the black fabric pants with dark blue coat, the blue trench coat with brown bags in the first row in column (i), the sky blue down jacket with orange snow boots in the second row in column (i), and addressing the bleeding issues in the fabric components (e.g., consistency pattern in the bag component in the fifth row of column (i). It is evident that in the region without specific fabric swatch reference, our method, driven by the incorporation of design priors, is capable of generating more creative and innovative fashion designs as shown in column (I).

Table 3 Quantitative ablation comparisons for module.

Method	SSIM↑	PSNR↑	FID↓	LPIPS↓	MOS↑
Baseline	0.87	19.30	43.34	0.13	3.32
Baseline + SCR	0.87	19.32	40.38	**0.12**	3.64
Baseline + SCR + SCT	**0.90**	**22.49**	**34.83**	**0.12**	**3.92**

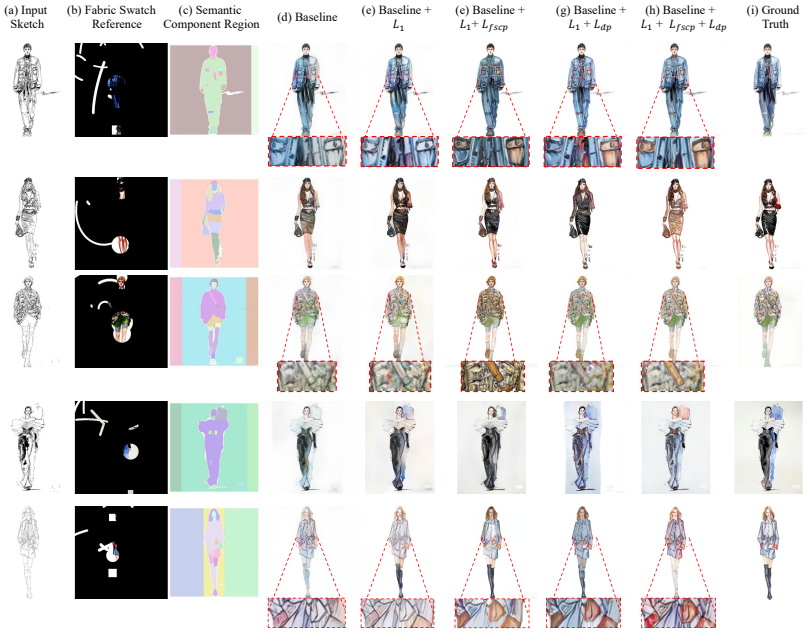

Fig. 7 Visual ablation comparisons for loss function. Please zoom in to view the details.

Automatic sketch rendering method comparison. Figure 4 provides visual qualitative comparisons with state-of-the-art automatic sketch rendering methods. AlacGAN [16] often exhibits component pattern bleeding, especially noticeable in background and hair regions in column (c). CWR [31], despite its implicit region information integration, tends to apply similar color patterns across sketches in column (d). ControlNet [9] and UniControl[10] depend heavily on the quality of its text-to-image model and text prompts. This reliance can lead to recurring color schemes and backgrounds that don't align well with input sketches, as illustrated in (e) where facial details in generated images deviate from the original sketches. Both (f) and (g) are results fine-tuned using the LoRA; (f) is fine-tuned with our SketchCouture dataset, while (g) is fine-tuned on the publicly available fashion clothing illustration dataset. These methods struggle to preserve the structural information of the sketches, with severe deformations occurring in the head and hands, and the background appearing dull. Our method showcases well-coordinated and diverse color styles across different examples and successfully mitigates the component pattern bleeding issue owing to the semantic region perception capability.

Table 4 Quantitative ablation comparisons for loss function.

Method	SSIM↑	PSNR↑	FID↓	LPIPS↓	MOS↑
Baseline	0.87	19.77	41.77	0.13	2.02
Baseline + $\mathcal{L}_1$	0.88	22.39	36.96	0.12	3.26
Baseline + $\mathcal{L}_1$+ $\mathcal{L}_{fscp}$	0.89	22.49	35.40	0.12	3.84
Baseline + $\mathcal{L}_1$ + $\mathcal{L}_{fdp}$	<u>0.89</u>	<u>22.45</u>	**31.69**	0.12	<u>3.90</u>
Baseline + $\mathcal{L}_1$ + $\mathcal{L}_{fdp}$ + $\mathcal{L}_{fscp}$	**0.90**	**22.49**	<u>34.83</u>	**0.12**	**3.92**

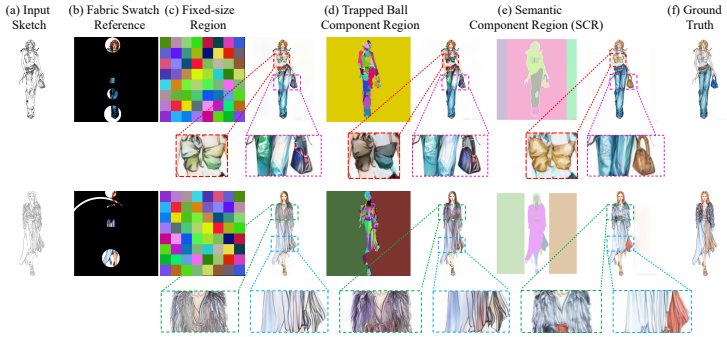

Fig. 8 Visual comparisons of different segment types.

4.4.2 Quantitative evaluation

Table 1 showcases our method's superior performance on IDB in terms of PSNR and FID scores, indicating a closer resemblance to real color images compared to other Reference-based sketch rendering techniques. Furthermore, our method also achieves the best MOS scores, further confirming that our results are in line with human visual perception as well as significantly fewer artifacts. It is worth noting that although PITI slightly outperforms our method in SSIM. One underlying reason is that PITI greatly preserves the color information rather than the texture of the fabric sample references as shown in the last row in column (h) of Figure 5, resulting in high brightness and contrast in generated images, which are good for SSIM. The other reason is that PITI does not incorporate fashion design priors to form more creative color matching. The Creativity implies a certain difference in color and texture from the original image, hence the SSIM score of our method is a little lower than that of PITI. The coloring results in the second and fourth rows of Figure 7 clearly demonstrate the creativity of our method. PITI's lower PSNR score, resulting from its overemphasis on brightness and contrast, also hampers the preservation of essential sketch details, as evidenced by line distortions and loss of fine sketch elements. Table 2 further demonstrates the superiority of our method.

4.5 Ablation Study

Module ablation comparison. We conducted qualitative and quantitative assessments on variants of our model to validate the effectiveness of each module. As depicted in Figure 6, column (d), the baseline UNet model without guidance shows significant pattern bleeding. Incorporating the Semantic Component Region (SCR) reduces this

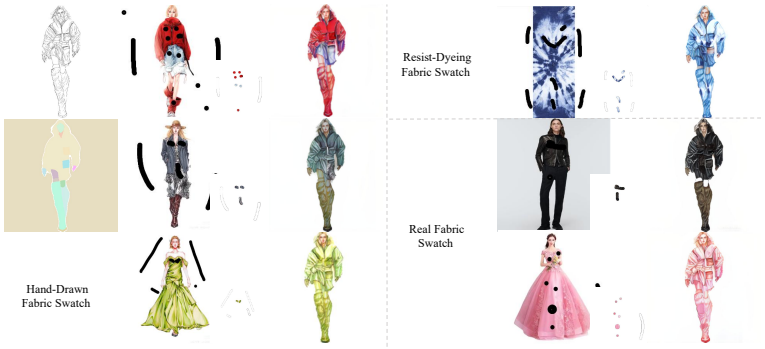

Fig. 9 Visual comparisons of different fabric swatch reference. Please zoom in to view the details.

issue (see column (c)), with notable improvements in FID and MOS scores (refer to Table 3, second row). However, due to a lack of semantic understanding across different components, the generated components with the same semantics do not have consistent color patterns(column (e), first row), where the texture of the pants on the thighs and calves is inconsistent. The implementation of the SCT block with a multi-head attention mechanism effectively resolves this, demonstrated by the consistent color patterns in the pants (column (f), first row). Our method, as summarized in the last row of Table 3, achieves the best scores across all metrics, substantiating its overall superiority.

Loss function ablation comparison. Figure 7 presents the qualitative results of the ablation study on the loss function. The baseline, employing only Adversarial loss, exhibits visual artifacts in the generated images. Specifically, as shown in the first, third, and fifth rows, gradient colors and patterns are bleeding in the collar, background, and waist. Additionally, the clothing in the third row appears blurry, resulting in low-quality visual effects. The integration of Pixel-Wise loss $\mathcal{L}_1$ enhances image clarity but introduces unrealistic color combinations in clothing, evident in the first row, where the color combination is simple. The addition of FDP loss $\mathcal{L}_{fscp}$ improves component independence, the third row of column (f) demonstrates the backpack strap is a creative and coordinated bright yellow color, but some issues like unrealistic color combinations in the denim jacket (column (e), first row) persist. The addition of FDP loss $\mathcal{L}_{fdp}$ improves color combinations, guiding images towards a multimodal distribution aligned with fabric swatch references, some issues like semantic color inconsistency in the pockets (column (g), first row) persist. After adding FDP loss $\mathcal{L}_{fdp}$ and FSCP loss $\mathcal{L}_{fscp}$, the colors of the two pockets remain consistent (column (h), first row), and remains the component independence the third and fourth row of column (h). Furthermore, in the fourth row of column (h), the background of the character's shoulder creates a harmonious composition while highlighting the character. Quantitative analysis in Table 4 reveals that each added loss function incrementally optimizes our method, with the FDP loss producing the lowest FID score. These results underscore the distinct effectiveness and importance of each loss function in our study.

Different types of segment region map comparison. Figure 8 illustrates the impact of various region map types on image generation, highlighting the superiority

Table 5 Ablation comparisons for different segment types.

Method	SSIM↑	PSNR↑	FID↓	LPIPS↓	MOS↑
32 × 32 Fixed-size regions	0.87	19.12	43.06	0.13	3.02
Trapped Ball region maps [40]	0.89	21.52	37.42	0.12	3.70
Semantic component region [34]	**0.90**	**22.49**	**34.83**	0.12	**3.92**

Fig. 10 Visual comparisons of different datasets. Additional results are provided in the appendix 15.

of the Semantic Component Region (SCR) approach. Our experiments utilized region maps of fixed 32 × 32, Trapped Ball region maps, and SCRs, all of which process image content regionally but vary in the granularity of this processing. For the fashion dataset, image perception guided by regions containing rich and reasonable semantics is crucial. The 32 × 32 fixed-size region lacks semantic consistency within the same component region since it does not consider the segmentation of clothing components, as presented in column (c). Using a Trapped Ball region map may result in overly meticulous segmentation [40]. Especially when fashion sketch is complex, it might hamper the learning of the style of each region and the implicit semantic relationships between them. Specifically, it can predict the same color for adjacent regions that represent different semantics, as shown in the position of the long skirt in the second row of column (d). In contrast, due to the strong comprehension capability of the large-scale segmentation method [34], offers more accurate component regions. This is evident in column (e), where components like the handbag, and long skirt in column (e) have no color pattern bleeding issues and achieve compatibility colors within the components. The effectiveness of the SCR approach is further substantiated by the data presented in Table 5, which performs relatively well in all evaluation metrics.

Different types of fabric swatch reference. Figure 9 illustrates the effect of various types of fabric swatch references on sketch rendering. Our experiments utilized references from hand-drawn fabric swatches, real resist-dyeing fabric swatches, and real cloth fabric swatches. It is worth noting that our approach consistently achieves reasonable rendering results across all these reference types.

Generalizability Across Different Datasets. Given the difficulties in verifying the authenticity of sketches in real-world fashion datasets, we chose to assess the generalizability of our approach through the hand-drawn anime dataset [41]. As shown in Figure 10, our DPG-SCT effectively addresses the problem of color bleeding in anime image rendering. More comparison results are displayed in the appendices.

5 Conclusion

In this study, we introduce the Design Priors Guided Semantic Component Transformer (DPG-SCT) Network, a novel framework for the rendering of hand-drawn fashion sketches. Although numerous efforts have integrated semantic priors to guide

the synthesis from sketches to images; our work uniquely leverages the segmentation capabilities of a large-scale semantic segmentation model. More importantly, the utilization of semantic component regions provided by the large model forms the core of our investigation. On one hand, we employ Semantic Component Transformer (SCT) blocks to facilitate semantic-region-aware perception. These blocks enable local semantic perception within individual components and contextual semantic understanding across different components, thus ensuring not only the overall color aesthetics but also the distinctiveness of each component with hand-drawn fabric patterns. On the other hand, we propose Fashion Semantic Component Prior (FSCP) losses to ensure regional pattern style consistency in the rendered images. Through these methodologies, we effectively leverage the semantic component regions provided by the large model, thereby enhancing the performance of fashion rendering. Additionally, we propose a Fashion Design Prior loss function, specifically designed for our hand-drawn sketch rendering task. This Loss constrains the generated color distribution to adhere to a multimodal color distribution, thereby preventing excessive color diversity and ensuring compliance with the fundamental principles of fashion design. Our experiment demonstrates that the output fashion illustrations not only meet optimal measurement criteria but also enhance creative visual expression for designers. Additionally, we contribute a hand-drawn fashion dataset SketchCouture and delineate a methodology to emulate hand-drawn fashion sketch styles.

6 Limitation

Theoretically, our pipeline can also be applied to other sketch2image tasks. However, our work has been focused on fashion sketches, and domain knowledge of fashion design such as "multimodal distribution", and "semantic-region-aware" has been exploited to devise fashion design prior loss. For other tasks, such as face sketch images, incorporating geometric priors of faces would be helpful.

Acknowledgments

This work was supported by the National Natural Science Foundation of China(NSFC) under Grants NO.61976038, NO.61932020, NO.61772108.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ning Wang, Anqi Zou, Shuge Qian, and Baoli Sun. The first draft of the manuscript was written by Ning Wang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

7 Appendices

In the supplementary material, we present additional experimental results, including datasets, visualizations of the color distribution of generated images, further

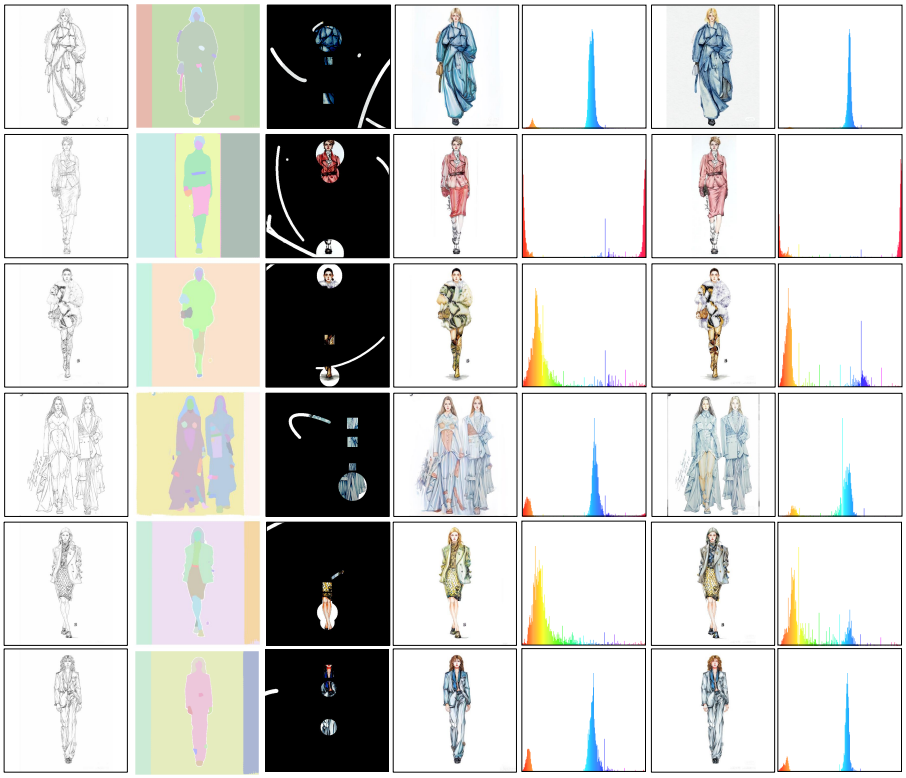

Fig. 11 The color distribution map of our results. (a) Input sketch, (b) Semantic Component Region, (c) Fabric reference, (d)(f) Our results, (e)(g) Color distribution map.

comparative and higher-resolution experimental results, and extended experiments demonstrating the combination of diffusion-based methods with our approach.

Figure 11 demonstrates that the color distribution of our rendering results exhibits a multimodal nature, which substantiates the effectiveness of our fashion design prior loss. Furthermore, our method displays a pronounced regional character, such as the handbags in the first row and the clothing areas in the fourth row, underscoring the efficacy of incorporating semantic component regions.

Figure 12 shows the details of our SketchCouture dataset. It includes fashion illustration, histograms of the hue-color distribution for each fashion illustration, two different types of sketches, as well as semantic component region maps obtained from various semantic segmentation models. SketchCouture features a diverse collection of hand-drawn fashion illustrations and encompasses a wide array of clothing types, including fashion wear, streetwear, and formal dresses. Besides, it employs various mediums such as markers, watercolors, and oil paintings, highlighting the dataset's diversity. Besides, different garment components (such as tops, pants, etc.) exhibit unique textures, and the overall outfit needs to satisfy aesthetic coordination. Importantly, it can be observed that the histogram distributions of the dataset exhibit a multimodal color distribution nature (multiple peaks in its color composition), demonstrating the uniqueness of the dataset. This aligns with the fundamental principles of

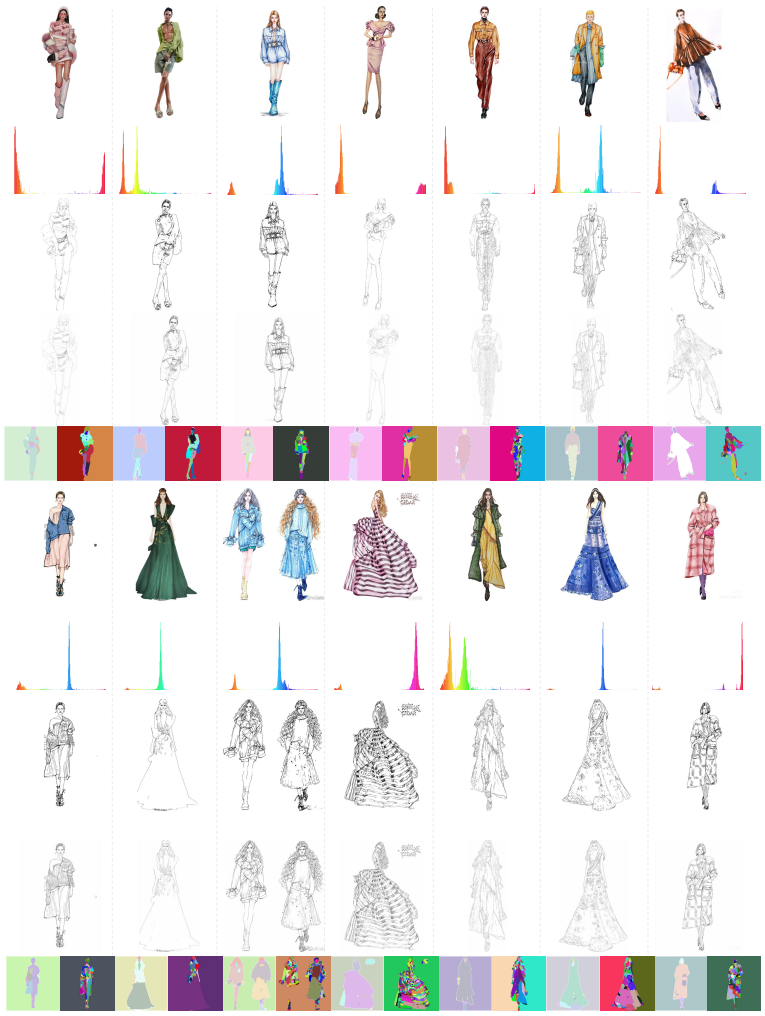

Fig. 12 The SketchCouture dataset provides examples of fashion illustrations, each accompanied by histograms of the hue color distribution and two types of sketches. Additionally, two types of semantic component region maps are available [34, 40].

fashion coordination, which suggest that in addition to the basic colors of black, white, and gray, the number of colors or textures in an outfit should ideally not exceed three [42, 43]. In summary, in the task of fashion sketch rendering, several challenges arise that need to be carefully managed to produce high-quality rendering results. These challenges encompass three main aspects: variability, complexity in clothing coordination, and multimodal color distribution.

Figure 13 presents additional visual comparisons with SCFT [1], SGA [6], DPGAN [7], TextureGAN [4], PITI [8]. These comparisons clearly demonstrate the superior quality of our results.

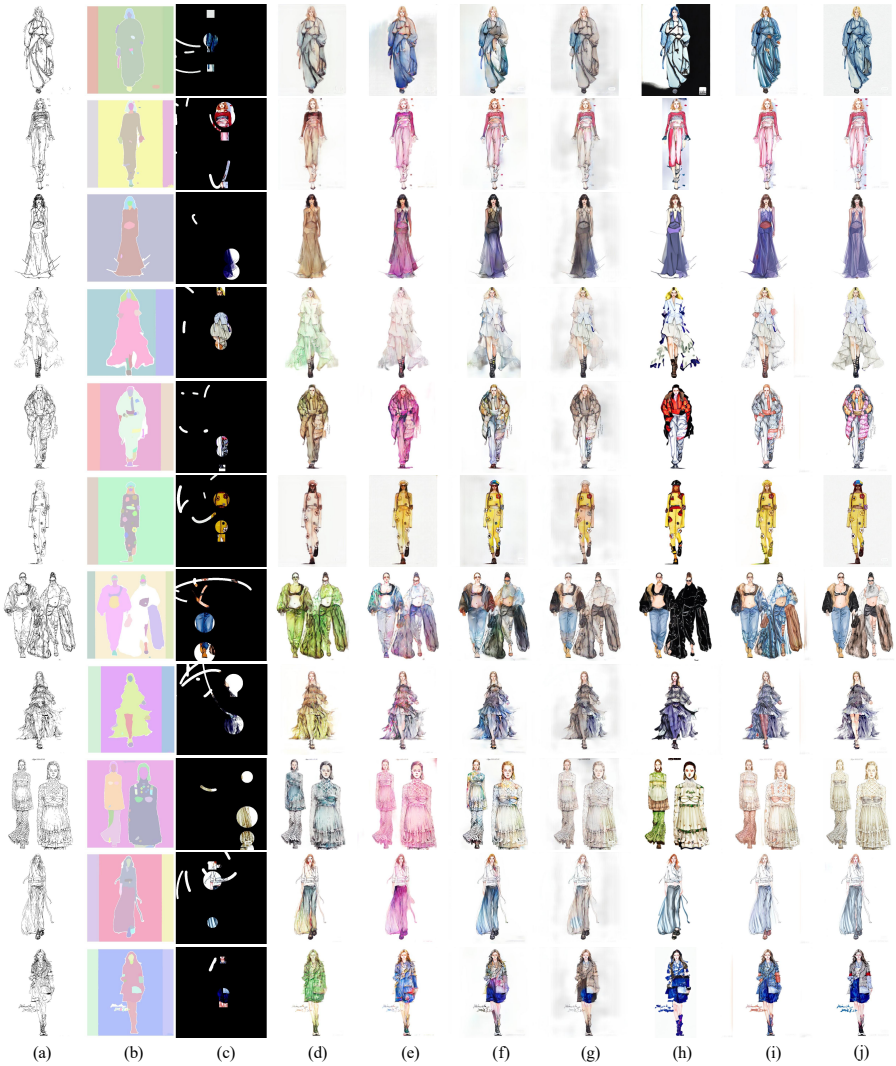

Fig. 13 Additional visual comparisons with state-of-the-art reference-based sketch rendering methods. (a) Input sketch, (b) Semantic component region, (c) Fabric reference, (d) SCFT [1], (e) SGA [6], (f) DPGAN [7], (g) TextureGAN [4], (h) PITI [8], (i) Ours, (j) Ground Truth.

Figure 14 further exhibits an array of high-resolution results generated by our model. We also provide a display of the diversity in coloring results for the same image, achieved through various fabric swatch references.

Figure 15 shows the generalizability in the anime domain of our method. Considering the distinct characteristics between hand-drawn anime and fashion, particularly the need for rich color in anime[7], we have eliminated the fashion prior loss. This adjustment reduces the influence of multimodal color distribution constraints. As shown in figure15, our DPG-SCT significantly reduces color bleeding issues compared to the best anime generation diffusion models ControlNet [9]. Our method maintains consistent colors within the same semantic

Fig. 14 High-resolution results of our model.

regions and preserves the structural information. This confirms that our model's semantic-region-aware perception capabilities are effectively adaptable to various sketch-rendering tasks.

To demonstrate the generalizability of our method, we also process 1,173 drafts from a Design-to-Real task [44]. As shown in the left of Figure 16, we used a user-friendly input method to obtain the fabric swatch from the reference image, (c) is manually selected from the reference on the left side of (c) using the Gradio drawing tool. Our method enables accurate semantic-region-aware perception, allowing it to generate aesthetically pleasing color combinations, such as nude-colored shoes and bags.

Fig. 15 A comparison of automatic anime sketch coloring methods without text prompts: the second column displays the generated results of fine-tuned ControlNet [9] using the anime dataset collected from [41]. The third column shows the generated results using the publicly available ControlNet. The fourth column presents the results of our DPG-SCT.

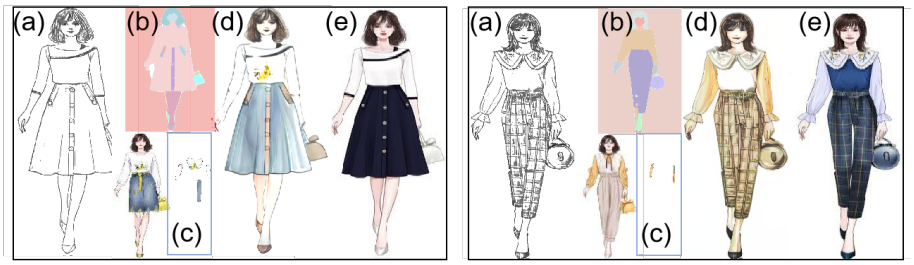

Fig. 16 (a) Input Sketch, (b) Semantic Component Region. (c) Fabric Swatch Reference. (d) Ours. (e) Ground Truth.

References

[1] Lee, J., Kim, E., Lee, Y., Kim, D., Chang, J., Choo, J.: Reference-based

sketch image colorization using augmented-self reference and dense semantic correspondence. In: Computer Vision and Pattern Recognition, pp. 5801–5810 (2020)

[2] Wang, N., Niu, M., Dou, Z., Wang, Z., Wang, Z., Ming, Z., Liu, B., Li, H.: Coloring anime line art videos with transformation region enhancement network. Pattern Recognition **141**, 109562 (2023)

[3] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. arXiv **abs/1406.2661** (2014)

[4] Xian, W., Sangkloy, P., Agrawal, V., Raj, A., Lu, J., Fang, C., Yu, F., Hays, J.: Texturegan: Controlling deep image synthesis with texture patches. In: Computer Vision and Pattern Recognition, pp. 8456–8465 (2018)

[5] Yan, H., Zhang, H., Liu, L., Zhou, D., Xu, X., Zhang, Z., Yan, S.: Toward intelligent design: An ai-based fashion designer using generative adversarial networks aided by sketch and rendering generators. IEEE Transactions on Multimedia **25**, 2323–2338 (2023)

[6] Li, Z., Geng, Z., Kang, Z., Chen, W., Yang, Y.: Eliminating gradient conflict in reference-based line-art colorization. In: European Conference on Computer Vision, pp. 579–596 (2022)

[7] Dou, Z., Wang, N., Li, B., Wang, Z., Li, H., Liu, B.: Dual color space guided sketch colorization. IEEE Transactions on Image Processing **30**, 7292–7304 (2021)

[8] Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arVix:abs/2205.12952 (2022)

[9] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: International Conference on Computer Vision, pp. 3836–3847 (2023)

[10] Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., Wong, K.-Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)

[11] Yang, L., Pang, K., Zhang, H., Song, Y.-Z.: Annotation-free human sketch quality assessment. International Journal of Computer Vision, 1–22 (2024)

[12] Qi, Y., Su, G., Wang, Q., Yang, J., Pang, K., Song, Y.-Z.: Generative sketch healing. International Journal of Computer Vision **130**(8), 2006–2021 (2022)

[13] Cui, Y.R., Liu, Q., Gao, C.Y., Su, Z.: Fashiongan: display your fashion design

using conditional generative adversarial nets. In: Computer Graphics Forum, vol. 37, pp. 109–119 (2018)

[14] Wu, D., Yu, Z., Ma, N., Jiang, J., Wang, Y., Zhou, G., Deng, H., Li, Y.: Styleme: Towards intelligent fashion generation with designer style. In: Human Factors in Computing Systems, pp. 1–16 (2023)

[15] Furusawa, C., Hiroshiba, K., Ogaki, K., Odagiri, Y.: Comicolorization: semi-automatic manga colorization. In: Special Interest Group for Computer GRAPHICS, 2017, pp. 12–1124 (2017)

[16] Ci, Y., Ma, X., Wang, Z., Li, H., Luo, Z.: User-guided deep anime line art colorization with conditional adversarial networks. In: International Conference on Multimedia, pp. 1536–1544 (2018)

[17] Zhang, L., Li, C., Simo-Serra, E., Ji, Y., Wong, T.-T., Liu, C.: User-guided line art flat filling with split filling mechanism. In: Computer Vision and Pattern Recognition, pp. 9889–9898 (2021)

[18] Maejima, A., Kubo, H., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Graph matching based anime colorization with multiple references. In: Special Interest Group for Computer GRAPHICS, 2019 (2019)

[19] Wang, N., Niu, M., Wang, Z., Hu, K., Liu, B., Wang, Z., Li, H.: Region assisted sketch colorization. IEEE Transactions on Image Processing **32**, 6142–6154 (2023)

[20] Sun, T.-H., Lai, C.-H., Wong, S.-K., Wang, Y.-S.: Adversarial colorization of icons based on contour and color conditions. In: International Conference on Multimedia, pp. 683–691 (2019)

[21] Li, Y.-k., Lien, Y.-H., Wang, Y.-S.: Style-structure disentangled features and normalizing flows for diverse icon colorization. In: Conference on Computer Vision and Pattern Recognition, pp. 11244–11253 (2022)

[22] Wu, S., Yang, Y., Xu, S., Liu, W., Yan, X., Zhang, S.: Flexicon: Flexible icon colorization via guided images and palettes. In: International Conference on Multimedia, pp. 8662–8673 (2023)

[23] Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition, pp. 5967–5976 (2017)

[24] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)

[25] Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback. In: European Conference on Computer Vision, pp. 129–147 (2025)

[26] Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. In: Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2017)

[27] Wang, Z., Wang, X., Xie, L., Qi, Z., Shan, Y., Wang, W., Luo, P.: Styleadapter: A unified stylized image generation model. International Journal of Computer Vision, 1–18 (2024)

[28] Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision, 1–20 (2024)

[29] Jian, J., Yang, S., Xinyang, Z., Zhenyong, F., Jian, Y.: Unicanvas: Affordance-aware unified real image editing via customized text-to-image generation. International Journal of Computer Vision, 1573–1405 (2025)

[30] Sheng, B., Li, P., Gao, C., Ma, K.-L.: Deep neural representation guided face sketch synthesis. IEEE Transactions on Visualization and Computer Graphics **25**, 3216–3230 (2018)

[31] Cao, R., Mo, H., Gao, C.: Line art colorization based on explicit region segmentation. In: Computer Graphics Forum, vol. 40, pp. 1–10 (2021)

[32] Cho, Y., Lee, J., Yang, S., Kim, J., Park, Y., Lee, H., Khan, M.A., Kim, D., Choo, J.: Guiding users to where to give color hints for efficient interactive sketch colorization via unsupervised region prioritization. In: Conference on Computer Vision and Pattern Recognition, pp. 1818–1827 (2023)

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing systems **30** (2017)

[34] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: International Conference on Computer Vision, pp. 4015–4026 (2023)

[35] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)

[36] Wang, X., Xie, C. Liangbinand Dong, Ying, S.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops, 2021, pp. 1905–1914 (2021)

[37] Chan, C., Durand, F., Isola, P.: Learning to generate line drawings that convey geometry and semantics. In: Computer Vision and Pattern Recognition, pp. 7905–7915 (2022)

[38] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

[39] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision Workshops, pp. 0–0 (2018)

[40] Zhang, S.-H., Chen, T., Zhang, Y.-F., Hu, S.-M., Martin, R.R.: Vectorizing cartoon animations. IEEE Transactions on Visualization and Computer Graphics **15**(4), 618–629 (2009)

[41] Anonymous, community, D., Branwen, G.: Danbooru2020: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset (2021)

[42] Steels, L., Belpaeme, T., *et al.*: Coordinating perceptually grounded categories through language: A case study for colour. Behavioral and brain sciences **28**(4), 469–488 (2005)

[43] Singh, S.: Impact of color on marketing. Management decision **44**(6), 783–789 (2006)

[44] Han, Y., Yang, S., Wang, W., Liu, J.: From design draft to real attire: Unaligned fashion image translation. In: International Conference on Multimedia, pp. 1533–1541 (2020)