Original article

Textile Research Journal

Textile Research Journal
0(0) 1–18
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00405175251356681
journals.sagepub.com/home/trj
Sage

iRetexturing: Intelligent fashion items retexturing via diffusion models

Yarui Zhang[1], Yao Jin[1,2] and Xin Huang[1]

Abstract

To meet the fashion industry's increasing demand for intelligent tools capable of manipulating and retexturing fashion items while maintaining structural integrity, an area where existing methods often lack photorealistic quality and intuitive control, we introduce iRetexturing, a novel framework that integrates diffusion models and geometric priors to tackle these challenges. The iRetexturing method utilizes a three-stage pipeline: high-resolution preprocessing with masked super-resolution and semantic segmentation, quadripartite texture synthesis using grid-based tiling and boundary-aware regeneration, and ControlNet-guided diffusion with dual spatial constraints (Canny edges and depth maps), incorporating innovations such as parametric texture modulation, partial repainting with seam refinement, and real-time adaptability across diverse materials. Evaluations conducted on 4400 fashion images reveal that iRetexturing outperforms state-of-the-art methods including diffuseIT, achieving a learned perceptual image patch similarity of 0.1385 and structural similarity of 0.8323 compared with diffuseIT's 0.1618 and 0.8074, respectively, despite diffuseIT's lower Fréchet inception distance (54.96 versus 75.25), highlighting iRetexturing's superiority in fine-grained texture replacement and high-fidelity textile design. By combining diffusion-based generation with geometric priors, iRetexturing enables precise manipulation of surface characteristics, bridging the gap between conceptual prototypes and production-ready assets, and offering transformative potential for the fashion industry by streamlining the design-to-production process and fostering creative innovation.

Keywords

Retexturing, diffusion models, repainting, quadripartite continuous, geometric preservation, adaptive conditioning.

When browsing online, customers frequently find themselves contemplating the aesthetic appeal of fashion items adorned with various textures. They imagine how different fabrics, patterns, and finishes would transform the look and feel of a garment or accessory. Similarly, when designing fashion products, designers are often required to edit and change textures multiple times to achieve the perfect appearance that aligns with their vision and meets market demands. Therefore, it is of utmost importance to develop advanced tools that can realistically retexture fashion items with ease. These tools not only enhance the virtual shopping experience by providing customers with a more immersive and interactive environment, but also boost design efficiency by streamlining the texture editing process. By ensuring visual consistency through precise texture adjustments, such technology enables designers to create cohesive and visually appealing collections. Moreover, the impact of this technology extends far beyond the fashion industry. It also has broad applications in various fields, including retail, film, entertainment, gaming, and virtual reality. Retailers can leverage these tools to offer augmented reality (AR)-based virtual try-ons, allowing customers to visualize how different textures and styles would look on them without physically trying on the fashion items. Filmmakers, on the other hand, can seamlessly alter costumes during postproduction, achieving the desired visual effects with greater flexibility and precision.

Existing techniques of fashion item retexturing[1–4] have demonstrated significant progress in image

[1] Zhejiang Sci-Tech University, School Computer Science and Technology (School of Artificial Intelligence), Hangzhou, China
[2] Zhejiang Provincial Innovation Center of Advanced Textile Technology, Shaoxing, China

Corresponding authors:
Yao Jin, Zhejiang Sci-Tech University, School Computer Science and Technology (School of Artificial Intelligence), Hangzhou, 310018, China.
Email: jinyao@zstu.edu.cn

Xin Huang, Zhejiang Sci-Tech University, School Computer Science and Technology (School of Artificial Intelligence), Hangzhou, 310018, China.
Email: xhuang@zstu.edu.cn

generation quality and design flexibility. Recent approaches[5–7] have enhanced physical plausibility through 3D surface normal prediction for ultraviolet (UV) map guidance. Some works[8,9] have investigated texture transfer paradigms for material replacement, while others[10,11] have explored virtual garment synthesis through data-driven approaches that enhance texture resolution from base geometric templates. Notably, the emergence of diffusion models[12–15] has demonstrated remarkable potential for generalized texture manipulation in fashion applications.

Despite these advancements, several critical limitations remain unaddressed in current fashion item retexturing techniques. The frequent failure to preserve geometric fidelity, as observed in earlier methods,[1–4] often results in visual artifacts such as distortions in geometric patterns or surface irregularities. Recent 3D-based approaches[5–7] are constrained by computationally intensive reconstruction pipelines and fall short in achieving fine-grained details and intuitive user interactions. In addition, generative approaches, including diffusion models,[12–15] face challenges in maintaining coherent local feature consistency, particularly in complex multiobject scenarios where contextual relationships and spatial dependencies require precise preservation. These shortcomings highlight a gap in achieving both realism and efficiency, limit practical applicability for designers and consumers, and underscore an underexplored challenge in ensuring seamless and contextually coherent texture manipulation. Collectively, they indicate a need for innovative solutions that balance visual quality, computational feasibility, and usability in fashion design and related fields.

To address these challenges, we leverage the robust image repainting capabilities of diffusion models and innovatively repurpose them for fashion item retexturing, integrating auxiliary neural networks and dual input streams of texture and fashion item images to optimize the repainting process with enhanced realism and robustness. By incorporating quadripartite continuity into our diffusion-based framework, our method ensures seamless, tileable textures that eliminate boundary artifacts and maintain visual coherence across expansive fabric surfaces, offering a theoretically grounded advancement over existing approaches for fashion design and texture-intensive applications. As demonstrated in Figure 1, our approach exhibits exceptional adaptability in managing geometrically intricate textures (e.g., woven architectures) while preserving fine accessory details, thereby addressing the growing demands of mass customization and creative fashion design. The second row exemplifies our novel texture expansion strategy, where localized texture repainting achieves seamless large-scale pattern alignment on fashion items.

Our principal contributions are threefold.

1. We propose a diffusion model-driven pipeline for fashion item retexturing that ensures seamless

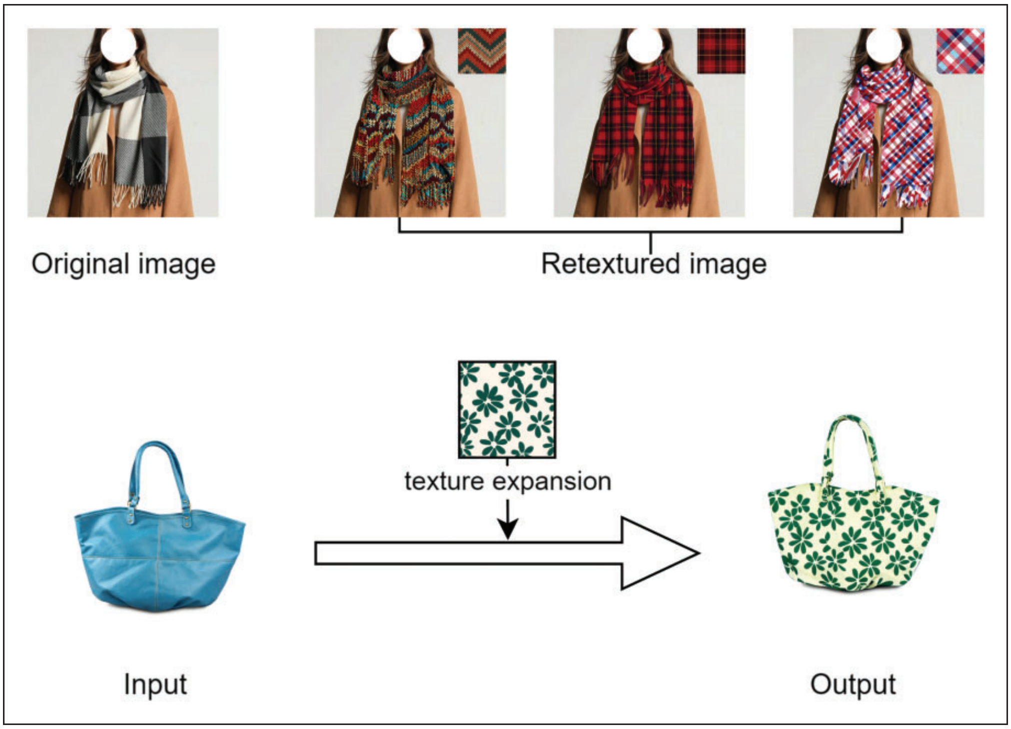

Figure 1. Our results of retexturing fashion items.

visual integration of single-source materials while enabling targeted texture customization.

2. We present a novel partial repainting technique that imbues textures with quadripartite continuity through precise seam refinement, guaranteeing omnidirectional pattern coherence.
3. We introduce a parametric retexturing system that enables precise visual modulation by manipulating tile repetition frequency, offering designers granular control over surface pattern density and orientation.

Related work

Fashion items generation

Recent advancements in artificial intelligence (AI)-driven fashion synthesis have been propelled by two competing technological frameworks: generative adversarial networks (GANs)[16] and diffusion models.[17–19] While both demonstrate substantial progress in design automation and consumer-oriented customization, their distinct methodological approaches present unique advantages and limitations.

GAN-based systems have revolutionized fashion design automation through three principal innovations. First, the CTS-GAN framework[12] achieves attribute disentanglement by decoupling color, texture, and shape into independent latent spaces, enabling non-experts to generate professional-grade designs with a 23% improvement in Style Diversity Index. Second, the Cross-domain Feature Fusion Framework[13] bridges creative intent and technical realization through sketch-to-garment translation, demonstrating 37.6% higher operational efficiency than conventional design methods. Third, DFDGAN[14] addresses multicategory compatibility via specialized architectural design, synthesizing stylistically coherent ensembles (e.g., top–bottom combinations) while reducing Fréchet inception distance (FID) by 15.2%. Notably, Quantum-GAN[15] pioneers computational optimization through feature reduction and evolutionary strategies, achieving a 92% qubit requirement reduction without compromising output fidelity.

In contrast, diffusion models exhibit distinct advantages through probabilistic denoising processes. TexControl[20] employs a two-stage pipeline to convert sketches into detailed 2D garment renderings with optimized texture resolution. DiffFashion[21] implements structure-aware appearance transfer for 2D clothing items, though its direct retexturing efficacy remains constrained. Karagoz et al.[22] demonstrated semantic-driven textile pattern synthesis, albeit with limitations in meeting specific commercial design requirements.

Diffusion-based architectures demonstrate theoretical superiority over GAN frameworks through three synergistic mechanisms: the stochastic differential equation framework inherently supports multistage refinement processes that expand output diversity while mitigating mode collapse; a self-correcting sampling architecture enables precise visual attribute manipulation through natural language conditioning or spatial guidance masks without requiring generator restructuring; and the model's nonadversarial gradient paradigm combines deterministic convergence properties with scalable parameter optimization, effectively resolving GANs' chronic limitations in architectural rigidity and training instability. This tripartite advantage establishes a new paradigm for controllable image synthesis that reconciles generation quality with computational efficiency through its unique blend of progressive refinement and mathematical tractability.

Building upon these advancements, our research harnesses diffusion models' intrinsic repainting capacity for fashion item retexturing. Through the strategic application of repainting algorithms on texture-overlay images, we achieve simultaneous preservation of structural integrity and innovative texture integration. This methodology not only enhances output realism but also demonstrates improved robustness compared with existing diffusion-based generation techniques.

The quality of raw materials, such as yarn and fabric, plays a pivotal role in the fashion item generation pipeline, directly affecting the aesthetic and functional attributes of synthesized products. Recent AI-driven advancements have expanded beyond generative techniques to enhance quality control in textile manufacturing. For instance, Hu et al.[23] introduced a predictive framework for yarn quality fluctuations, leveraging a multicorrelation parameter feature subspace mechanism. By combining kernel principal component analysis (KPCA) with a deep belief network (DBN) optimized via particle swarm optimization (PSO), their method achieves precise yarn quality forecasting, laying a robust foundation for downstream fashion synthesis. Complementarily, Hu et al.[24] proposed a U-Net-based approach for detecting sewing breaks in fabrics, attaining a detection accuracy of 95.75% through stitch contour uniformity analysis. This technique ensures fabric integrity, a critical prerequisite for high-quality fashion outputs. Together, these studies underscore the synergy between AI applications in textile quality assurance and fashion item generation, highlighting the transformative potential of deep learning across the textile-to-fashion continuum.

Fashion item retexturing

The inherent complexity of fashion item retexturing has led to the development of two principal methodological approaches. The first paradigm utilizes 3D reconstruction and rendering techniques to achieve precise geometric fidelity-preserving accuracy while replicating intricate garment details.[6,7,25] The second strategy employs direct texture mapping through neural estimation of dense UV maps, effectively circumventing 3D reconstruction requirements.[3,5,26,27]

3D reconstruction-based methodologies remain prevalent in industrial applications due to their exceptional capacity for capturing complex geometric dependencies. These techniques facilitate high-fidelity UV mapping and 3D deformation through sophisticated mesh reconstruction pipelines while inherently accommodating complex fabric distortion patterns characteristic of flexible materials. Notably, He et al.[25] demonstrated the efficacy of this paradigm through their single-image texture transfer framework, which successfully preserved critical geometric features, including fold structures and shading variations. However, despite achieving photorealistic rendering quality, the substantial computational complexity inherent in these reconstruction pipelines presents formidable obstacles for real-time implementation scenarios.

Direct UV mapping strategies adopt an alternative pathway by leveraging neural networks to establish texture correspondences directly from input images. These approaches, however, exhibit critical limitations in geometric fidelity due to incomplete surface modeling, often resulting in spatial inconsistencies during advanced editing operations. Jafarian et al.[5] enhanced this paradigm through self-supervised 3D normal prediction, generating temporally stable UV maps without full geometric reconstruction. Despite achieving improved physical plausibility in real-world scenarios, the method demonstrates restricted editability when manipulating texture boundaries or angular feature relationships, with edge artifacts persisting due to insufficient 3D structural comprehension.

Our methodology introduces a novel synthesis of Canny edge detection for surface effect analysis and depth estimation algorithms for geometric detail modeling. Experimental validation confirms that this combinatorial approach, when coupled with diffusion models, substantially enhances retexturing quality relative to conventional techniques. The proposed method effectively reconciles the precision limitations of direct UV mapping with the computational burdens of 3D reconstruction, establishing an optimized equilibrium between photorealism and operational efficiency.

Method

Overview

This study focuses on retexturing fashion item imagery by replacing surface textures while preserving the original geometric details of the items. To achieve this, we proposed an intelligent retexturing approach (iRetexturing). It employs techniques including super-resolution, semantic segmentation, and image tiling to generate optimized inputs for diffusion model-based repainting.

As shown in Figure 2, our intelligent retexturing framework employs diffusion models to jointly process fashion item images and textures through three core computational phases. The “(a) Image enhancement and preprocessing” stage generates high-quality visual inputs via super-resolution refinement and semantic segmentation, producing precise masked regions for targeted editing. Building upon this foundation, the “(b) Texture expansion” module ensures continuity by adaptively tiling input textures to match item geometry while preserving structural coherence. These processes are further regulated by “(c) ControlNet guidance,” which derives spatial constraints from input images to adjust texture synthesis parameters. The framework culminates in the SDXL (Stable Diffusion XL) model,[28] executing partial repainting through dual conditioning mechanisms that combine textual prompts with spatial constraints, selectively modifying masked fashion item regions while maintaining physically consistent blending between new textures and preserved structural features.

Image enhancement and preprocessing

As illustrated in Figure 2(a), the preprocessing of images of fashion items prioritizes enhanced data quality to achieve well-defined semantic segmentation results. To obtain precise mask areas for designated fashion items and guarantee segmentation accuracy, we first address the fundamental requirement of improving input image resolution. This quality enhancement proves critical because high-definition images preserve essential visual details that support improved boundary detection and semantic interpretation. For this resolution refinement process, we applied SDXL[28] repainting for image super-resolution, a technique that successfully restores fine-grained textures and structural elements in low-quality input images.

When processing low-resolution images of fashion items, super-resolution becomes essential for quality enhancement. First, the original low-resolution image I_{low} is scaled to the base resolution res_{HD} through partitioned processing with $n \times n$ chunks. By default, res_{HD} is set to 1024×1024 pt with $n \times n$ chunks, aligning with the standard size required by SDXL. The upscaled image is then divided into multiple overlapping subregions $\{I_i\}_{i=1}^{n}$, each of which maintains an overlap ratio r_{overlap}.

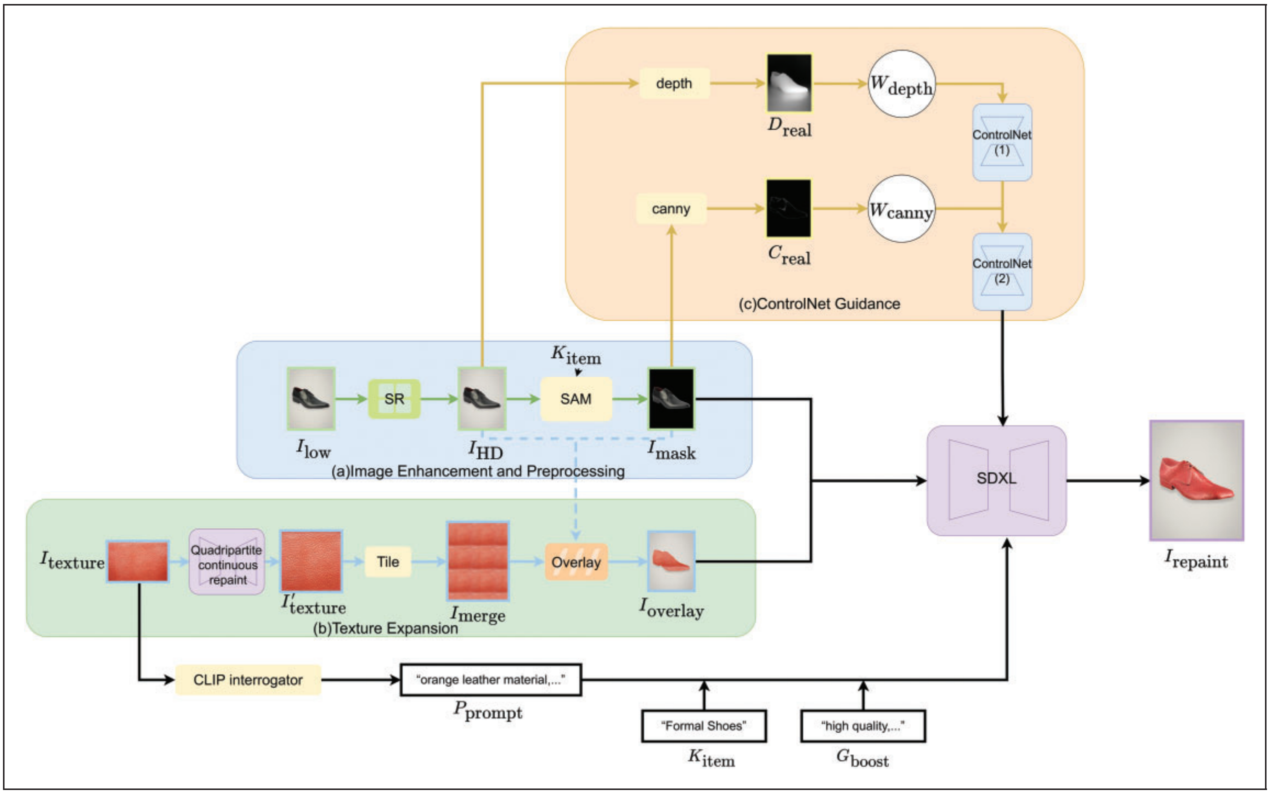

Figure 2. Our intelligent retexturing workflow of fashion items via diffusion models.

During the super-resolution task, we employ a conditional diffusion model for joint noise reduction and block reconstruction. Each low-resolution subregion I_i acts as the conditional input, guiding the generation of corresponding high-resolution outputs I_i' through reverse diffusion. This process is mathematically formulated as

$$p_\theta(\mathbf{x_{HR}}|\mathbf{x_{LR}}) = \int p_\theta(\mathbf{x_{HR}}|\mathbf{x_t})q(\mathbf{x_t}|\mathbf{x_{LR}})d\mathbf{x_t} \tag{1}$$

where $\mathbf{x_{LR}}$ represents the low-resolution image patch I_i. The conditional probability $p_\theta(\mathbf{x_{HR}}|\mathbf{x_t})$ models high-resolution reconstruction using the super-resolution network, whereas $q(\mathbf{x_t}|\mathbf{x_{LR}})$ characterizes the forward noise-addition process from $\mathbf{x_{LR}}$ to the intermediate state $\mathbf{x_t}$.

The generation workflow involves two phases: (1) progressive noise injection from $\mathbf{x_{LR}}$ to $\mathbf{x_t}$ through forward diffusion, followed by (2) iterative noise removal from $\mathbf{x_t}$ to $\mathbf{x_{HR}}$ final via reverse diffusion. Governed by Equation (1), this dual-phase mechanism ensures structural consistency with the original input while enhancing textural details, effectively addressing the inherent limitations of blur and detail loss in conventional super-resolution approaches.

The process culminates in seamless aggregation of all enhanced subregions $\{I_i'\}_{i=1}^{n}$ into a unified high-definition output I_{HD} at target resolution res_{HD}. Our algorithm strategically combines partitioned processing with intelligent stitching mechanisms to optimize computational efficiency for large-scale images. The conditional diffusion process effectively leverages low-resolution structural priors while replenishing high-frequency details, achieving fidelity-preserving super-resolution.

The enhanced I_{HD} undergoes semantic segmentation through the SAM[29] method using item-specific descriptors K_{item}, formally expressed as

$$(M_{\text{item}}, I_{\text{mask}}) = SAM(I_{\text{HD}}, K_{\text{item}}) \tag{2}$$

yielding two critical outputs: (1) the binary segmentation mask M_{item} pinpointing target fashion items, and (2) the alpha-channel composition I_{mask} with nonitem regions nullified (black background). These processed elements subsequently serve as primary inputs for downstream algorithmic components.

Texture expansion

Quadripartite continuity is a key geometric constraint in pattern design and texture synthesis, aimed at producing seamless, tileable patterns. This approach divides a design into four interrelated regions, ensuring that visible seam artifacts are eliminated when patterns are repeated. It enforces first-order continuity by aligning texture values at boundaries and can extend to higher-order continuity, such as quadratic continuity, which matches gradients for smooth transitions under scaling or deformation. This theoretical framework underpins the creation of visually cohesive patterns across various scales and transformations.

The significance of quadripartite continuity shines in fields including clothing, textiles, and image processing. In the fields of fashion pattern design, quadripartite continuity constitutes a critical geometric constraint for seamlessly tileable textures requiring expansive fabric coverage. This structural coherence requirement guarantees visual integrity during material application by eliminating perceptible boundary artifacts. In fabric design, it facilitates consistent patterns that remain intact despite stretching or folding. Our method enhances intrinsic pattern continuity through adaptive region repainting algorithms, enabling optimized replication efficiency while preserving semantic coherence with designated fashion item structural elements. Beyond fashion, quadripartite continuity is crucial in digital rendering, ensuring seamless textures across expansive surfaces. Recent research[30–32] highlights its pivotal role in advancing texture synthesis techniques. Extending to areas such as digital design, printed media, and material fabrication, quadripartite continuity establishes itself as an essential principle for seamless pattern generation across diverse industries.

As depicted in Figure 2(b), our two-phase enhancement process consists of: (1) quadripartite continuous texture refinement through partially repainting, and (2) adaptive tiling based on optimized quadripartite continuous texture. The processed texture achieves precise alignment with the HD fashion item's masked regions via geometric superposition, establishing structure–retentive texture integration that enables subsequent design modifications.

While achieving enhanced continuity, diffusion-synthesized textures frequently manifest boundary discontinuities (Figure 2(b), inset) typified by perceptible transition anomalies. To resolve these residual artifacts, we deploy SDXL's[28] constrained partial repainting technique with edge-aware operators, methodically resolving inter-region discontinuities while maintaining global textural consistency and chromatic uniformity.

Extending Zhang et al.'s[30] continuous tiling paradigm, our methodology differentiates itself through computational grid alignment protocols rather than latent space interventions. As detailed in Figure 3, the original texture map I_{texture} undergoes systematic grid-based partitioning and recombination through computational transformation pipelines. For nonisometric inputs, dimensional normalization to $L_{\text{square}} \times L_{\text{square}}$ is achieved via adaptive centroid cropping prior to tiling operations.

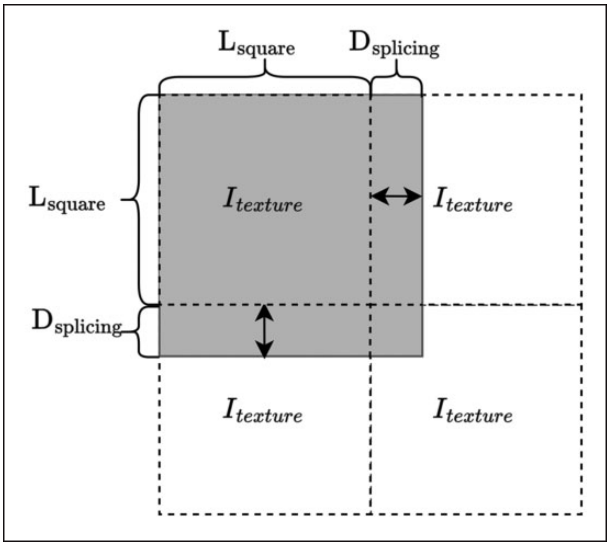

Figure 3. Illustration of the splicing and cropping of texture images.

The technical workflow comprises three stages: (1) composite 2 × 2 grid assembly from standardized texture units, (2) extraction of an expanded $L_{square}+D_{splicing}$-sized square containing strategic overlap regions, as shown in Figure 3 where the gray area represents the area after splicing and cropping, and (3) boundary-constrained regeneration of transitional areas. Through empirical validation, we establish optimal parameters as $L_{square} = 1024\,\text{px}$ (matching SDXL's[29] native resolution) with $D_{splicing} = 252\,\text{px}$ ($0.33 \times$ base resolution), achieving balanced continuity and detail preservation.

As depicted in Figure 4, our method leverages the IPAdapter[33] style transfer method (SDXL-based[28]) to perform localized regeneration of transitional zones in composite textures. This is accomplished through a blended latent diffusion process, which integrates latent diffusion with edge-aware denoising to iteratively refine latent representations under the guidance of a mask, thereby ensuring seamless quadripartite continuity. Figure 4 reveals how this approach systematically addresses boundary discontinuities through edge-aware denoising, establishing quadripartite continuity essential for texture tiling. The green highlighted regions in the figure emphasize the detailed improvements before and after the repainting process. Ultimately, this regeneration pipeline yields an enhanced texture, denoted as $I'_{texture}$.

Our texture repainting pipeline employs SDXL[28] diffusion models through systematic grid-based tiling operations. To establish foundational texture for model conditioning, we implement an optimized spatial

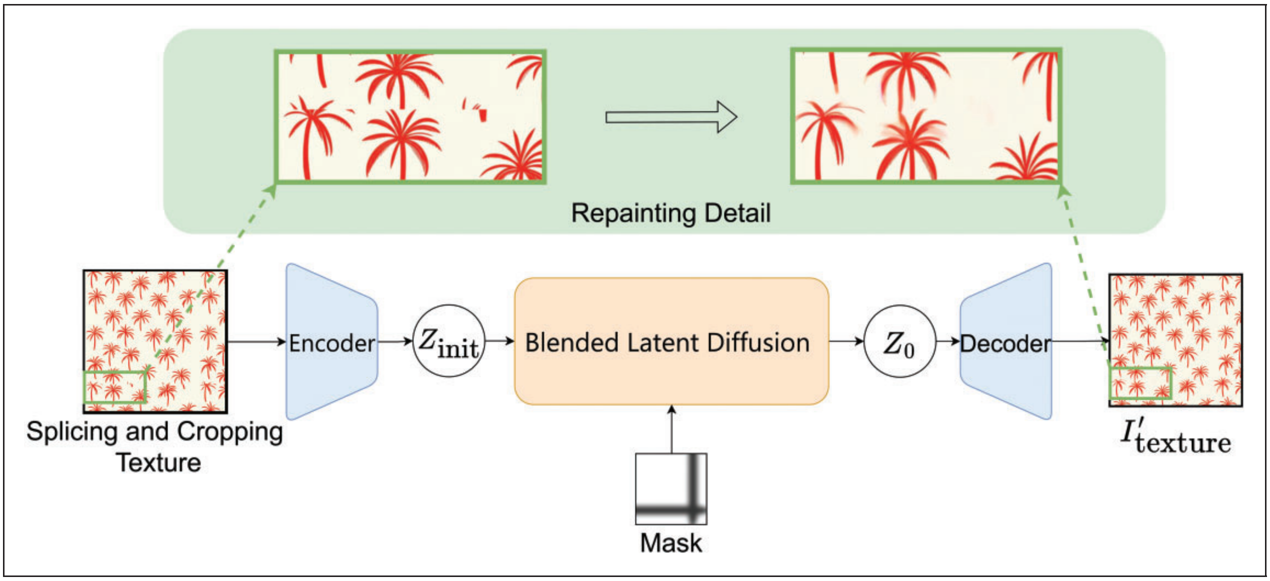

Figure 4. Architecture for synthesizing quadripartite-continuous textures.

repetition protocol: given input textures I_{ij} with dimensions $W \times H$, where indices $i, j \in \{0, 1, \ldots, n-1\}$ denote grid positions, our algorithm generates composite images I_{merge} through $n \times n$ matrix replication. This methodology provides dual technical benefits: (1) controlled upscaling through regular texture expansion, and (2) enhanced global continuity via systematic spatial repetition.

The grid assembly process preserves relative positioning through coordinate mapping $(x, y) \rightarrow (iW, jH)$, creating seamless transitions between adjacent tiles. As demonstrated in Figure 4, this approach enables production of arbitrarily large texture fields while maintaining local textural coherence: a critical requirement for subsequent diffusion-based regeneration processes. The resultant I_{merge} serves as both conditioning input and structural prior for SDXL's[28] denoising operations, effectively bridging discrete texture samples with continuous material synthesis.

To establish spatial correspondence between the tiled texture I_{merge} and the target fashion item I_{HD}, we implement a resolution matching protocol formally expressed as

$$T_{\text{processed}} = \Psi_{\text{crop}}(\Phi_{\text{scale}}(I_{\text{merge}}, \text{res}_{\text{HD}}), \text{res}_{\text{HD}}) \quad (3)$$

where Φ_{scale} implements adaptive geometric scaling that proportionally matches I_{merge}'s shortest edge to I_{HD}'s longest dimension, ensuring source texture coverage exceeds target areas. The subsequent Ψ_{crop} operator extracts res_{HD}-sized regions through center-aligned window sampling, achieving dimensional parity between texture and fashion item inputs. This preprocessing serves dual purposes: generation of geometrically compatible feature spaces for diffusion-based manipulation and preservation of texture resolution parity essential for high-frequency detail transfer. The method then implements alpha-channel compositing:

$$I_{\text{overlay}} = \Omega_{\text{mask}}(I_{\text{HD}}, T_{\text{processed}}, I_{\text{mask}}) \quad (4)$$

where Ω_{mask} performs mask-guided channel-wise composition, creating a hybrid input that encodes both the geometry of fashion items and texture details. This composite feature map provides structured initialization for SDXL's[28] latent space conditioning, enabling stable texture transfer through subsequent diffusion iterations.

Retexturing

To generate fashion item images with photorealistic textures while preserving geometric fidelity, we employ ControlNet[34] to govern the texture synthesis process. As depicted in Figure 2(c), the method processes two critical inputs: the masked image of the fashion item I_{mask} and the original image I_{real}. The pipeline first extracts structural guidance through Canny edge detection on I_{mask}, generating edge feature maps C_{real}, while simultaneously deriving spatial constraints via depth prediction from I_{real} to produce depth maps D_{real}. These complementary feature representations are then channeled into specialized ControlNet[34] branches for conditional synthesis.

The architecture implements weight specialization across distinct ControlNet[34] modules, with the edge-processing branch optimized through parameter adjustments w_{Canny} for contour preservation, and the depth-analysis branch calibrated via w_{depth} for three-dimensional consistency. This dual-branch conditioning mechanism enables the diffusion model to perform spatially coherent retexturing while maintaining the original item's structural integrity, ultimately achieving high-fidelity retexturing through geometrically constrained synthesis.

During the SDXL-based repainting phase,[28] the original texture $I_{texture}$ is processed through the CLIP interrogator to generate descriptive prompts P_{prompt} encapsulating textural characteristics. These semantic prompts are subsequently combined with categorical keywords K_{item} (specifying fashion item types) and enhancement directives G_{boost} (optimizing repainting intensity) through a structured text conditioning pipeline. The method integrates these textual controls with spatial constraints from ControlNet,[34] enabling SDXL[28] to execute geometrically constrained repainting. As a diffusion architecture, SDXL[28] performs localized texture synthesis on the masked region M_{item} of the overlaid image I_{overlay}, preserving nontarget areas through latent space regularization:

$$c_{text} = \tau_{\text{Clip}}\left(K_{\text{item}} \oplus P_{\text{prompt}} \oplus G_{\text{boost}}\right) \tag{5}$$

$$z_{t-1} = z_t - \epsilon_\theta(z_t, t, c_{text}, \tau_{\text{ControlNet}} \\ (w_{\text{Canny}} C_{\text{real}} + w_{\text{depth}} D_{\text{real}})) \tag{6}$$

where ϵ_θ denotes the central U-Net architecture in SDXL. The model receives noisy latent representations z_t at timestep t, augmented by textual conditioning via $\tau_{\text{ControlNet}}$ and spatial constraints from ControlNet.[34] This synthesis yields the retouched output I_{repaint}.

Notably, the VAE autoencoder's inherent information compression necessitates mask-guided compositing: we strategically overlay I_{repaint} onto the original low-resolution image I_{low} using mask region M_{item}, thereby preserving contextual elements while introducing enhanced textural details.

Experiments

Dataset

Our evaluation employs the Fashion Product Images Dataset,[35] containing 44,000 professionally captured high-definition fashion item images. This dataset was selected because it contains a variety of high-definition fashion items, making it suitable for retexturing testing. From this collection, we curated 4400 representative samples spanning 14 distinct apparel categories to ensure categorical diversity.

Evaluation metric

To establish comprehensive quantitative validation of our methodology, we employed three established image quality metrics from the computational photography and computer vision domains: learned perceptual image patch similarity (LPIPS),[36] structural similarity index (SSIM),[37] and Fréchet Inception Distance (FID).[38] LPIPS quantifies perceptual similarity through deep feature correlation analysis in pretrained convolutional neural networks. SSIM quantifies structural preservation through multiscale luminance-contrast covariance analysis. FID assesses generation realism through Fréchet distance computation in deep feature space between synthesized and reference distributions. These metrics collectively constitute a multidimensional assessment framework, rigorously evaluating texture reproduction fidelity, geometric–structural consistency, and photorealistic authenticity in our retexturing outputs.

Implementation details

The experimental setup utilizes NVIDIA GeForce RTX 4090 and RTX 3060 GPUs under Windows 11 with Python 3.10.8 and CUDA 11.8 acceleration for evaluating the baseline and our proposed methods, as detailed in the performance comparison (Table 3). Our implementation adopts SDXL[28] through ComfyUI's[39] modular architecture, enabling systematic hyperparameter optimization via its node-based computational graph interface. For multimodal conditioning, the controlnet-union-sdxl-1.0 model[40] integrates Canny edge detection and depth estimation modules with empirically optimized weighting coefficients (w_{Canny} =0.375, $w_{\text{depth}} = 1.0$) to balance structural preservation and spatial consistency.

In texture synthesis tasks, IPAdapter[33] maintained semantic guidance strength $\alpha = 1.0$ through cross-attention modulation, while SDXL's[28] denoising scheduler operated at $\varepsilon_{\text{base}} = 0.85$ for primary synthesis. All super-resolution and inpainting operations exclusively employed SDXL's[28] latent diffusion framework to maximize computational parallelism.

The texture extraction process for fashion items is conducted on a diverse set of images sourced from the Fashion Product Images Dataset,[35] encompassing various categories such as handbags, shoes, and garments. To initiate, the SAM[29] is employed to generate precise segmentation masks, isolating each fashion item from its background. Subsequently, the largest inscribed rectangle within each mask is computed to define the extractable texture region, discarding items where the rectangle is smaller than 256 × 256 pixels to ensure uniformity. Texture patches of 256 × 256 pixels are then randomly sampled from these regions, capturing diverse patterns, followed by Canny edge detection (thresholds 100 and 200) to filter out solid color patches, retaining only those with sufficient detail. This process, detailed in Table 1 and illustrated in Figure 5, is executed in batches with memory optimization, yielding a balanced texture dataset of thousands of patches suitable for retexturing applications in virtual fashion design.

Baseline methods

We conducted comprehensive comparisons with six state-of-the-art approaches across the domains of texture transfer and virtual try-on. In our taxonomy, we consider style transfer a subcategory of texture transfer, as both involve the manipulation of surface appearance while preserving structural or semantic content. Accordingly, the baseline for texture-oriented evaluation includes the following. *Texturize*[8] is an exemplar-guided synthesis framework employing bidirectional patch coherence optimization with physically based rendering (PBR) material matching, achieving photorealistic texture propagation through multi-scale feature alignment. *Texture-Reformer*[9] is a shape-consistent architecture combining view-specific UV texture completion with semantic graph convolutional networks, enforcing structural consistency through learnable geometric priors. *DiffFashion*[21] is a structure-aware diffusion framework for reference-based fashion design, which transfers garment appearance from a reference image while preserving the target clothing's shape and layout. The method explicitly models structure-aware correspondence to maintain spatial fidelity during style propagation. *DiffuseIT*[41] is a diffusion-based image translation approach that disentangles style and content representations. This disentanglement enables precise and controllable transfer of visual attributes, allowing for stylistic manipulation without compromising structural content: a capability particularly suited to fashion image synthesis.

Table 1. Pseudocode for texture extraction from fashion item images

```
Input: Fashion item image I
mask = SAM_segmentation(I) // Segment the fashion item from the image
mask = convert_to_binary(mask) // Convert mask to boolean array
rectangle = compute_largest_inscribed_rectangle(mask)
// Find largest rectangle within the mask
cropped_image = crop(I, rectangle) // Crop image to rectangle bounds
if (cropped_image.size < 256 × 256) then // Check if cropped region is too small
    return None
end if
// Exclude if rectangle is too small
texture_patch = random_crop(cropped_image, size=256)
if (not detect_edges(texture_patch, threshold1=100, threshold2=200)) then
// Apply edge detection to check texture content
    return None
end if
// Discard if no edges detected
return texture_patch
// Return the extracted texture patch
Output: texture_patch (256 × 256 texture patch, or None if extraction fails)
```

For virtual try-on tasks, the following models were adopted. *CP-VITON*[10] is a classical image-based virtual try-on method that employs thin-plate spline transformations with texture-preserving appearance fusion for geometric alignment. *Leffa*[11] is a pose-aware diffusion model developed by Meta AI researchers that preserves garment structural integrity during pose manipulation by implementing deformation field estimation techniques.

A comparative evaluation of these approaches elucidates their distinct mechanisms and interconnections. Exemplar-driven methods such as Texturize[8] and Texture-Reformer[9] excel in texture detail reproduction but falter in global structural alignment, as their

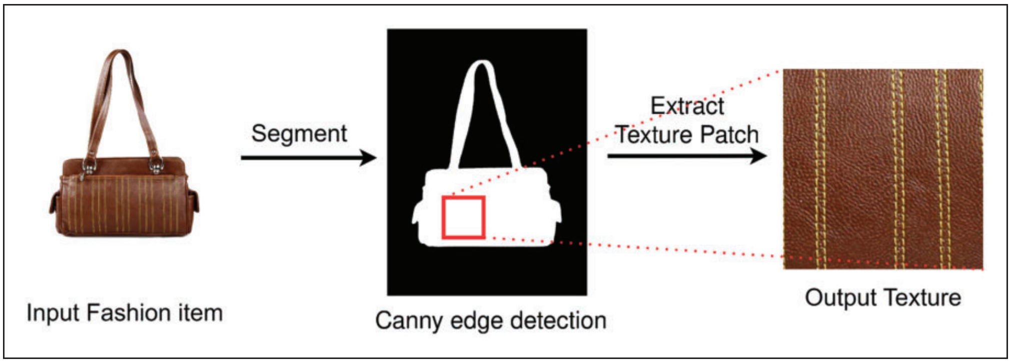

Figure 5. Schematic diagram of texture extraction process for a handbag.

formulations do not account for geometric variations. Virtual try-on techniques, such as CP-VITON[10] and Leffa,[11] adeptly manage pose-driven deformations yet compromise on texture precision due to their emphasis on spatial reconfiguration. Diffusion-based models, including DiffFashion[21] and DiffuseIT,[41] integrate these strengths by balancing texture and spatial optimization, though they differ in implementation: DiffFashion emphasizes appearance reconciliation via optimal transport, whereas DiffuseIT prioritizes feature disentanglement for style-content separation. Our proposed framework builds on these insights, leveraging diffusion models to achieve a synthesis that excels in both domains, as validated by experimental results showing enhanced photorealism and structural consistency across diverse scenarios.

Comparison with state-of-the-art methods

The evaluation protocol follows He et al.'s[25] paradigm, where input images serve as ground-truth references with identical texture applications, while our test dataset introduces critical enhancements: increased image resolution and enhanced structural complexity through multilayer fashion item compositions. These improvements explain metric variations compared to prior benchmarks.

Table 2 highlights that our proposed method demonstrates exceptional performance in fashion retexturing tasks, significantly enhancing key evaluation metrics across multiple dimensions. Specifically, it achieves an SSIM of 0.8323, reflecting substantial improvement in structural preservation compared with existing approaches. In addition, the LPIPS score of 0.1385 highlights its outstanding perceptual fidelity. While diffuseIT[41] records a lower FID of 54.96 compared with our 75.25, attributable to its style transfer approach emphasizing global feature alignment, our method excels in fine-grained texture replacement. It surpasses diffuseIT[41] with a notably lower LPIPS (0.1385 versus 0.1618) and a higher SSIM (0.8323 versus 0.8074), underscoring its superior balance of perceptual similarity and structural coherence. These findings affirm our framework's robust capability to generate structurally consistent and visually realistic retextured fashion items, establishing its advantage for high-fidelity applications.

Table 2. Comparison results of different methods

Method	LPIPS	SSIM	FID
Texturize[8]	0.5847	0.1204	277.5
Texture-Reformer[9]	0.6161	0.3792	274.9
CP-VTON[10]	0.1452	0.8043	97.91
Leffa[11]	0.2383	0.7432	99.42
DiffFashion[21]	0.3034	0.5690	260.8
DiffuseIT[41]	0.1618	0.8074	**54.96**
Our method	**0.1385**	**0.8323**	75.25

Our extended experimental analysis demonstrates that our framework robustly addresses a wide range of fashion items and imaging conditions by accurately rendering fine texture details and seamlessly adapting to variations in lighting and material properties. As evidenced in Figure 6 and further validated on a diverse set of web-sourced images, our method excels in reproducing complex textures, such as the pronounced wave pattern on the scarf, the sharp floral design on the backpack, and the lifelike leather finish on the boots, while effectively mitigating deficiencies observed in competing baselines. Specifically, in Figure 6, the first row (scarf) shows our method (column 9) integrating the blue wave pattern with superior natural shading and fabric fold alignment, unlike Texturize[8] (column 3), which applies the pattern uniformly but lacks depth, leading to structural deformations and chromatic shifts, and DiffFashion[21] (column 7), which improves color vibrancy but introduces light-induced inconsistencies. In the second row (backpack), our method (column 9) achieves flawless pattern alignment and vibrant colors with the floral texture, contrasting with CP-VTON[10] (column 5), which shows minor misalignment around the straps and noticeable texture detail erosion, and DiffuseIT[41] (column 8), which, despite refinement, over-smooths intricate patterns, stripping away essential fine details. In the third row

Table 3. Performance comparison of image processing methods across hardware

Method	Time/100 images (s)	Resolution	Pixels/image (MP)	Time/million pixels (s/MP)	Hardware
Texturize[8]	18.93	300 × 400	0.12	1.58	RTX3060
CP-VTON[10]	30.66	192 × 256	0.049	6.26	RTX3060
Texture-Reformer[9]	69.05	448 × 600	0.269	2.57	RTX3060
DiffuseIT[41]	2539.99	192 × 256	0.049	518.37	RTX3060
Leffa[11]	2335.17	900 × 1000	0.9	25.95	RTX4090
DiffFashion[21]	4412.74	256 × 256	0.065	672.0	RTX3060
Ours (RTX4090)	3494.30	768 × 1024	0.786	44.43	RTX4090
Ours (RTX3060)	12,885.30	768 × 1024	0.786	163.75	RTX3060

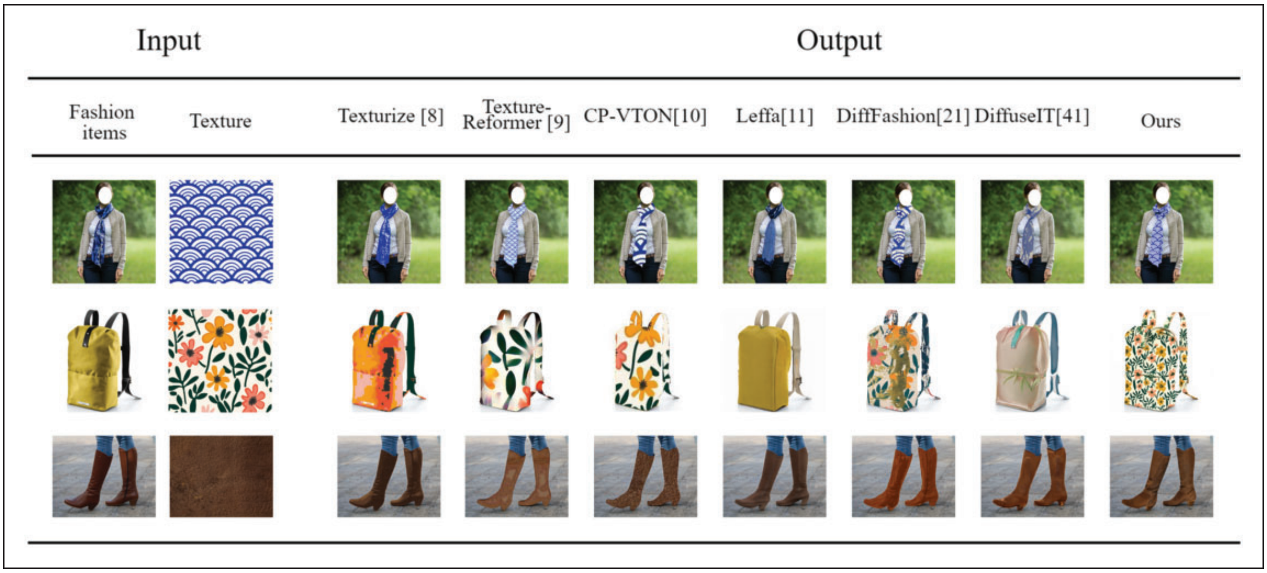

Figure 6. Comparison results of retexturing fashion items with different methods.

(boots), our method (column 9) delivers a lifelike leather texture with realistic wear and lighting, surpassing Texture-Reformer[9] (column 4), which suffers from significant edge degradation with smoother, less-detailed transitions, and Leffa[11] (column 6), which produces oversimplified, flatter textures lacking material authenticity. In contrast, our approach, by leveraging high-precision semantic segmentation masks, not only preserves both global structure and local texture fidelity under dynamic illumination conditions but also delivers a cohesive rendering that aligns closely with the inherent material properties of each fashion item. This detailed comparative evaluation underscores the clear advantage of our method in achieving superior texture precision, structural integrity, and adaptive performance across diverse experimental scenarios.

The superior performance of our method stems from its dual-channel conditioning mechanism, which establishes explicit texture–item correlations through targeted latent space manipulation. This enables three critical advantages: structural fidelity preservation through constrained diffusion sampling, high-frequency detail retention via texture-adaptive attention modulation, and chromatic consistency maintenance via color-aware guidance. Figure 7 demonstrates the versatility of our framework across a wider range of fashion items, showcasing its ability to handle different styles, textures, and complexities. This further highlights its robustness in diverse real-world scenarios.

Ablation study

Our ablation studies reveal that critical parameter configurations and architectural selections exert a deterministic influence on the generation quality. The method's auxiliary processing pipeline incorporates Canny edge detection and monocular depth estimation as dual preprocessing modules for ControlNet[34] conditioning. The Canny operator enhances texture alignment precision through contour fidelity preservation, while depth prediction enforces spatial coherence by maintaining 3D structural consistency with the original fashion item geometries. This synergistic preprocessing combination establishes foundational constraints that critically govern subsequent generation fidelity.

Our systematic evaluation revealed critical dependencies between the edge detection algorithms and the final synthesis quality. Figure 8 compares four Canny variants, Manga Lineart,[42] Lineart,[43] TEED,[44] and CannyEdge,[45] analyzing their texture processing capabilities under standardized experimental conditions.

The ablation study employed the optimized parameters of Depth Anything V2[46] on representative texture patches, with a fixed ControlNet[34] guidance weight at 0.675 across all configurations. For geometrically simple textures, Manga Lineart[42] and Lineart[43] achieved superior edge retention, whereas complex patterns revealed Lineart's[43] dominance in detail preservation. These findings establish Lineart[43] as the optimal general-purpose Canny implementation for our workflow, balancing structural coherence with textural fidelity across diverse fashion item textures.

Figure 9 systematically benchmarks three prominent depth estimation architectures: Depth Anything V2,[46] Zoe Depth,[47] and Depth Anything.[48] The experimental results demonstrate that Depth Anything V2[46] achieves superior surface detail reconstruction fidelity,

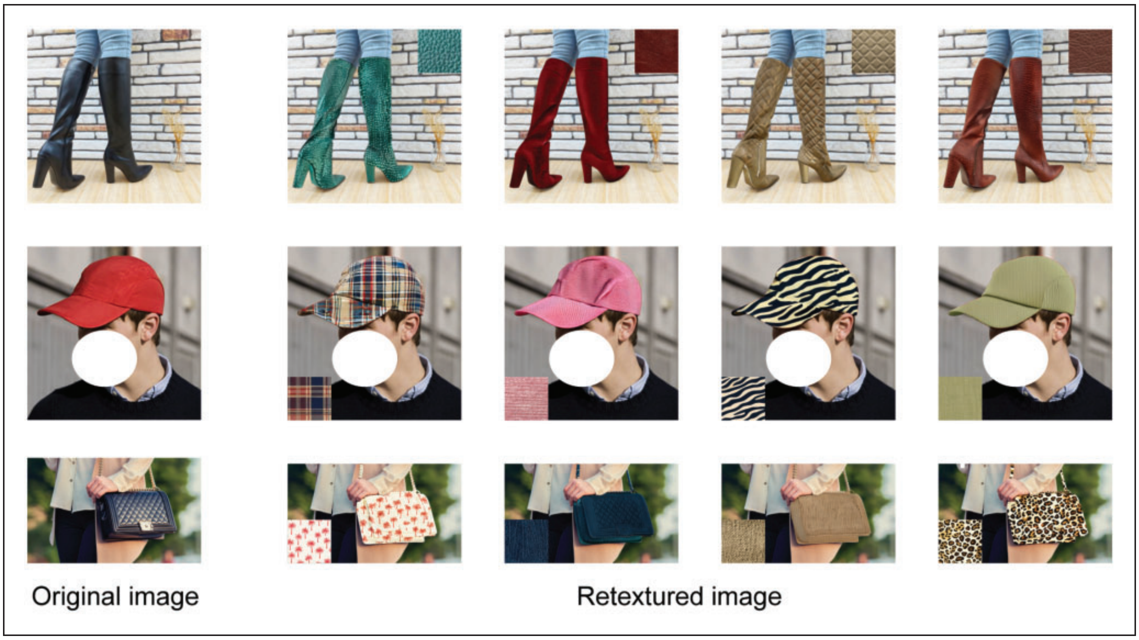

Figure 7. More results on retexturing fashion items.

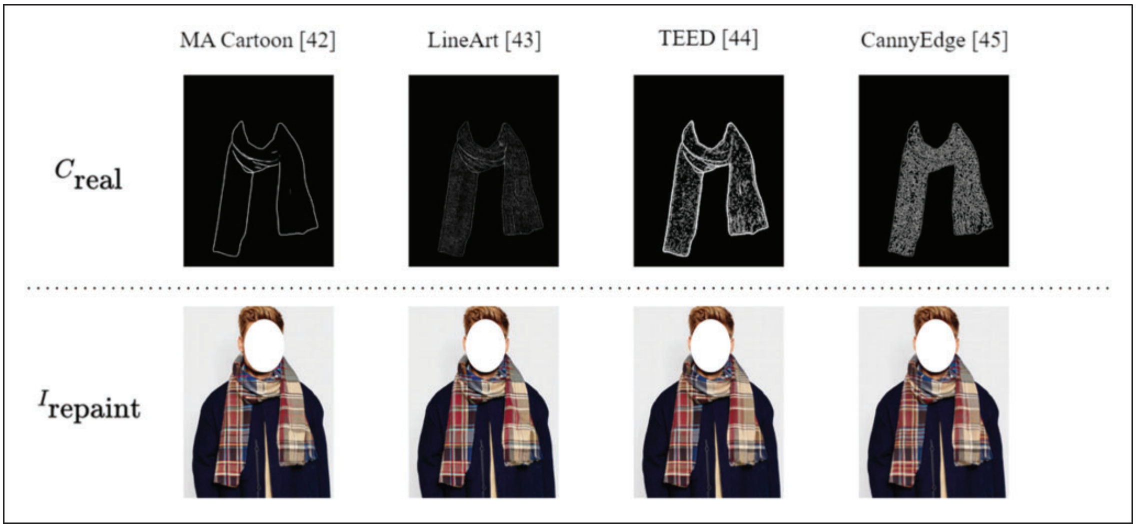

Figure 8. Comparison results with different Canny algorithms.

producing spatially coherent texture representations for fashion items. This superiority stems from its hybrid architecture, which combines multi-scale feature fusion with adaptive edge weighting. Depth Anything V2[46] exhibits cross-domain adaptability, maintaining robust performance across both geometrically simple and pattern-intensive textures.

Based on these observations, we adopted LineArt[43] for edge-aware texture alignment and Depth Anything V2[46] for spatial consistency as the optimal algorithmic pairing, offering an optimal balance between computational efficiency and reconstruction accuracy for most fashion item retexturing scenarios.

We systematically investigated the combined effects of the Canny edge detection weights (w_{Canny}) and depth perception weights (w_{depth}) on backpack texture generation through a comprehensive parameter space analysis, as shown in Figure 10. The results reveal that when w_{depth} falls below 0.25 (indicated by the blue frame), the generated images exhibit pronounced edge blurring, particularly in functional structures such as strap buckles and zipper tracks, accompanied by

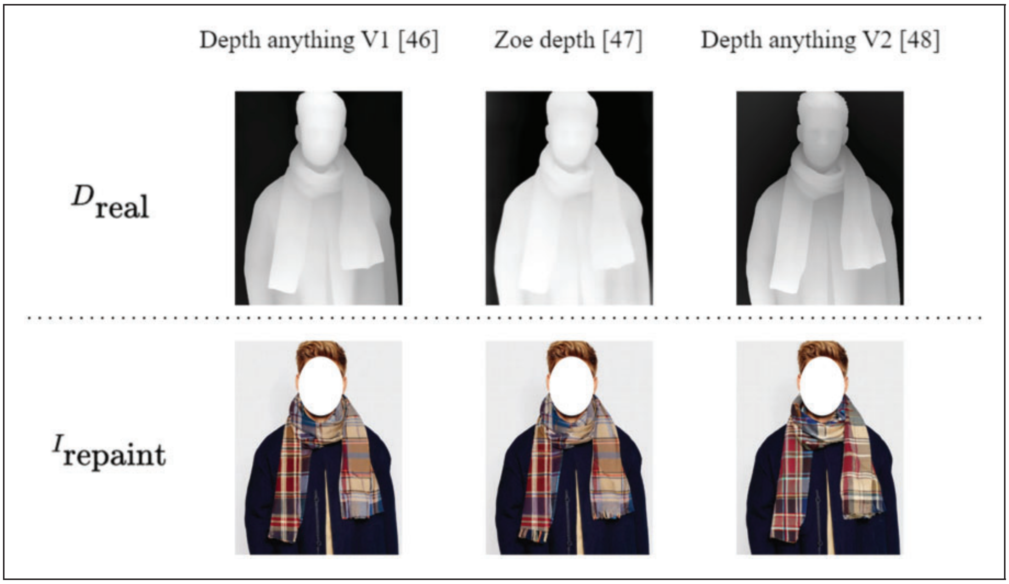

Figure 9. Comparison results with different depth estimation algorithms.

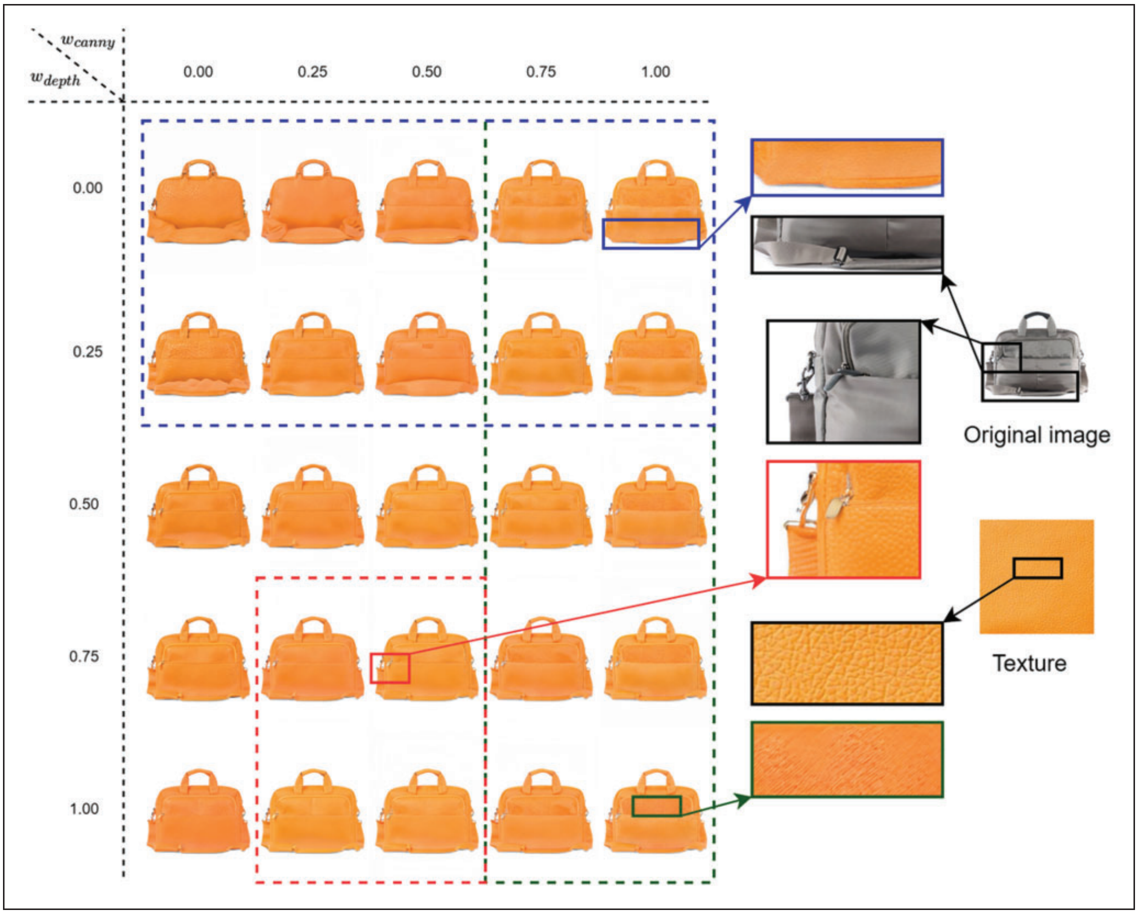

Figure 10. Sensitivity analysis of Canny and depth weights.

geometric distortions, as illustrated in the blue frame inset. Conversely, when w_{Canny} exceeds 0.50 (marked by the green frame), the texture semantic features degrade, manifesting as fragmented grid patterns and misaligned stripes, as shown in the green-framed inset. Empirically, the parameter range ($w_{\text{Canny}} = 0.25$–0.50 and $w_{\text{depth}} = 0.75$–1.00, delineated by the red frame) achieves an optimal balance between the edge constraint and the texture fidelity. As evidenced by the red-framed inset, this range yields high-quality results with precise zipper details and consistent texture patterns, ensuring accurate alignment of critical boundaries and natural adaptation of textures across complex surfaces. These findings provide essential guidelines for optimizing multimodal-controlled texture generation algorithms.

We investigated the mechanism by which the texture detail influences the replacement effects through a series of controlled experiments. As illustrated in Figure 11, three texture pairs were examined: (a) fabric leopard print versus (b) vector leopard pattern, (c) green leather versus (d) solid green, and (e) red fabric versus (f) solid red. Using leather shoes, baseball caps, and backpacks as benchmark apparel items, the SDXL model generated 18 visualization outcomes (6 object categories × 3 contrast groups). The experimental results indicate that texture images containing authentic material details (e.g., fabric leopard prints and leather textures) yield more 3D and realistic replacement effects compared to vector graphics or solid-color samples that lack such detail. Visual analyses show that textures with physical material characteristics (a, c, e) outperform their corresponding control groups (b, d, f) in terms of surface light field continuity (evident in the reflection areas on leather shoe vamp) and texture boundary sharpness (notably in the seams of cap brims).

Notably, we hypothesize that this phenomenon may be associated with the CLIP Interrogator's material semantic inference mechanism and the inherent characteristics of the SDXL[28] model's pretraining data distribution. Specifically, the high-frequency detail features present in authentic texture images may better align with CLIP's text-aligned material description paradigm, thereby enabling the diffusion model to generate more physically plausible texture mappings. This discovery offers critical insights for texture-guided prompt engineering methods, suggesting that future research could optimize prompt generation modules through the construction of a comprehensive fabric material lexicon.

A critical yet understudied aspect of fashion item retexturing lies in handling multiangular texture inputs, where surface orientation variations substantially alter perceptual coherence. We systematically evaluated the angular dependencies through controlled texture mapping experiments. As demonstrated in Figure 12, our method preserves textural continuity across 0–270° rotations while maintaining texture integrity through adaptive spatial warping, achieving visually consistent surface representations under varying projection angles.

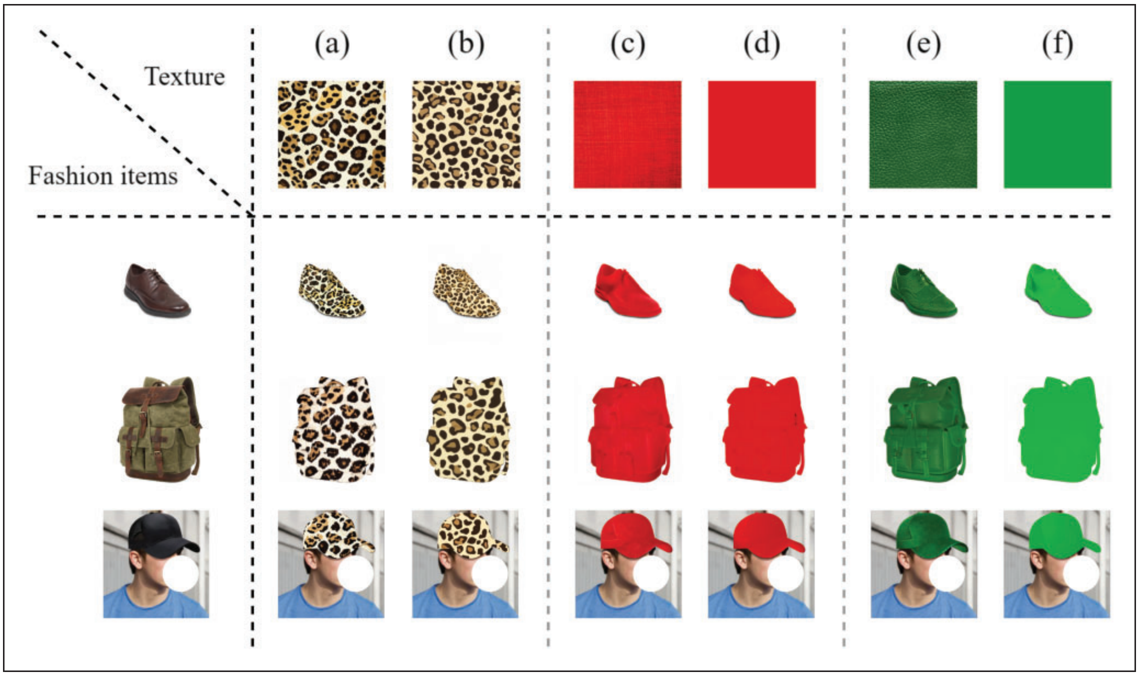

Figure 11. Texture effect on fashion retexturing results.

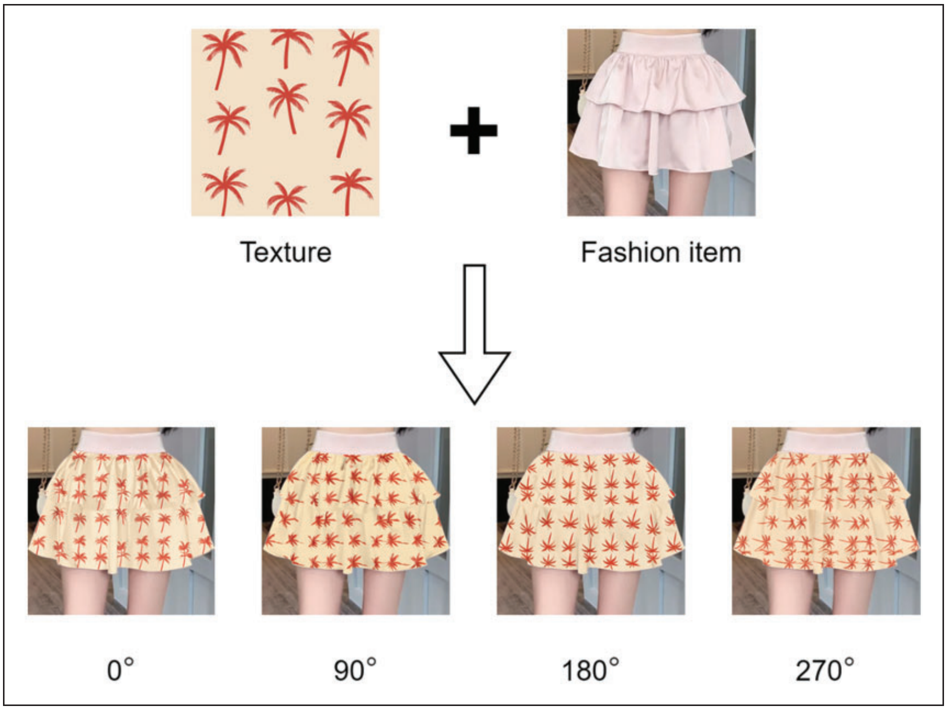

Figure 12. Retexturing results with different texture orientations.

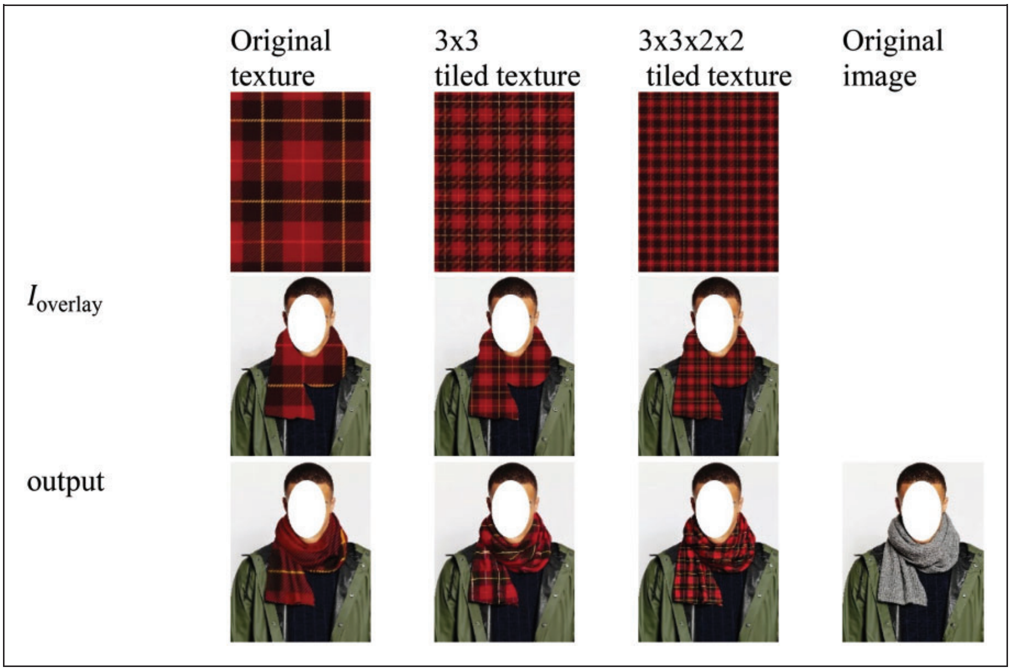

Figure 13. Retexturing results with different scaling sizes.

In addition, our method enables dynamic texture scaling through tile dimension editing. Figure 13 illustrates this capability using multiscale checkerboard inputs. The retextured outputs exhibit precise density preservation and maintain texture density consistency compared with the source textures. This scale-aware synthesis mechanism ensures geometric faithfulness across spatial resolutions, which is crucial for replicating both bold prints and intricate embroidery textures.

Our framework demonstrates robust parametric control over texture orientation and scaling operations, maintaining geometric and perceptual consistency across arbitrary angular variations (0–270°) and supporting *n*-times magnification while preserving texture fidelity, thereby establishing a solid foundation for future interactive editing systems with real-time adjustable parameters. These built-in orientation and scaling parameters establish a foundational control matrix that can be naturally extended to interactive design scenarios. By exposing these adjustment axes through a graphical interface, where designers could manipulate rotation sliders or scaling multipliers with real-time previews, the framework's core architecture already possesses the necessary mathematical scaffolding to support user-directed texture customization. Future implementations could further augment this parameterized system by introducing additional control variables for texture contrast adjustment, pattern phase shifting, and material property blending, thereby transforming the current automated pipeline into a versatile hybrid design tool that synergizes AI-driven generation with human creative guidance.

To ensure a fair evaluation of algorithmic time efficiency, we benchmarked multiple baseline methods on the RTX3060, using preprepared clothing masks as input (consistent with baseline requirements) and bypassing real-time mask generation via the SAM.[29] For equitable comparison across varying image resolutions, we introduced pixels/image (MP) and time/million pixels (s/MP) as standardized metrics. Our method demonstrates exceptional potential for practical deployment by efficiently processing high-resolution images (768 × 1024, 0.786 MP) across diverse hardware environments. On an RTX3060, it completes a batch of 100 images in 12,885 seconds (163.75 s/MP), making it accessible for resource-constrained settings such as small studios or research labs, while on an RTX4090, it achieves a faster processing time of 3494 seconds (44.43 s/MP), supporting high-performance applications such as real-time image generation in fashion or gaming. Compared with baselines, our method excels: Leffa[11] fails on the RTX3060 and requires an RTX4090 (25.95 s/MP, 0.9 MP), while low-resolution methods such as Texturize[8] (1.58 s/MP, 0.12 MP), CP-VTON[10] (6.26 s/MP, 0.049 MP), and Texture-Reformer[9] (2.57 s/MP, 0.269 MP) are inadequate for quality-demanding tasks. Less-efficient baselines, such as DiffuseIT[41] (518.37 s/MP, 0.049 MP) and DiffFashion[21] (672.0 s/MP, 0.065 MP), further underscore our method's superior balance of resolution, efficiency, and hardware adaptability, positioning it as a versatile solution for academic and industrial applications.

Computational profiling identified opportunities to enhance runtime efficiency, with the depth estimation module (9.4% of runtime), unmodified SDXL model (35%), and CLIP inversion process (13.0%) as primary bottlenecks. Targeted optimizations, such as lightweight depth estimation architectures, distilled SDXL variants with accelerated sampling, and compact text encoders or caching for CLIP inversion, could significantly improve the performance. These refinements require careful validation to maintain the perceptual quality and geometric accuracy. A phased approach, prioritizing depth model substitution and sampling acceleration, will enable the systematic evaluation of performance-quality tradeoffs, aligning with advancements in efficient generative architectures and reinforcing our method's scalability and deployment potential while ensuring high-fidelity outputs.

Conclusion and limitations

We have introduced a diffusion-based retexturing framework for fashion items that achieves realistic results by innovatively combining geometric texture constraints with generative refinement, enabling high-fidelity texture transfer while preserving structural integrity. This approach showcases the adaptive capacity of the diffusion models in managing the polymorphic inputs, markedly improving the spatial texture consistency. Experimental results underscore the critical role of parameter settings and architectural choices, particularly ControlNet's conditioning mechanisms, in determining the fidelity of the texture transfer, although performance varies with texture complexity, requiring tailored network adjustments for optimal outcomes. Looking ahead, future research will focus on systematically enhancing this methodology, with an emphasis on improving computational efficiency and optimizing the pipeline, such as by developing robust protocols for automated parameter tuning while upholding the precision demonstrated in our current implementation. Despite its limitations, this framework represents a significant advancement in fashion item retexturing and sets a strong foundation for further exploration.

Our diffusion-based retexturing framework for fashion items, while effective, encounters several notable limitations. Primarily, when the input texture material lacks realistic high-frequency details, the resulting texture replacements fail to deliver a convincing material appearance or pronounced three-dimensionality. This shortfall arises from the CLIP text-aligned material description paradigm, which struggles to capture the subtle nuances required for realistic texture rendering when dealing with less-detailed inputs, ultimately reducing the visual fidelity of the output. In addition, the method struggles when the input apparel is relatively flat and lacks inherent depth or layering, leading to generated outputs with diminished 3D and tactile qualities. This issue likely stems from the robustness limitations in the depth weighting mechanism of the ControlNet component, which impedes accurate depth cue representation. Furthermore, the absence of comprehensive 3D geometric reconstruction poses a challenge: fabric wrinkles require discontinuous texture rendering for physical accuracy, yet our approach produces continuous textures, resulting in physically inaccurate representations. This inaccuracy becomes particularly evident under challenging conditions, such as strong lighting, where the lack of geometric detail is more pronounced.

Acknowledgment

We sincerely thank Mr. Chen Li (ORCID: 0009-0007-5048-4217) for his valuable contributions to Section "Comparison with state-of-the-art methods" and "Ablation study" and 4.6 (Ablation Study) through providing essential experimental data support.

Data availability

The dataset analyzed in this study is publicly available on Kaggle under Creative Commons license. This dataset can be accessed and downloaded at: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset

The dataset is permanently archived with DOI: 10.34740/KAGGLE/DS/139630

All experimental implementations in this paper directly reference this dataset.[30]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Key R&D Programs of Zhejiang Province (grant number 2024SJCZX0026).

ORCID iDs

Yarui Zhang https://orcid.org/0009-0003-7924-8358
Yao Jin https://orcid.org/0000-0001-9518-7063
Xin Huang https://orcid.org/0000-0001-7113-5066

References

1. Shen J, Sun H, Mao X, Guo Y, and Jin X. Color-mood-aware clothing re-texturing. In: *2011 12th International Conference on Computer-Aided Design and Computer Graphics*. IEEE, 2011, pp. 151–154.
2. Xie Y, Mao H, Yao A, and Thuerey N. TemporalUV: Capturing loose clothing with temporally coherent UV coordinates." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 3450–3459.
3. Güler RA, Neverova N, and Kokkinos I. Densepose: Dense human pose estimation in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.
4. Ianina A, Sarafianos N, Xu Y, Rocco I, and Tung T. Bodymap: Learning full-body dense correspondence map. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13286–13295.
5. Jafarian Y, Wang TY, Ceylan D, et al. Normal-guided garment UV prediction for human re-texturing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4627–4636.
6. AlBahar B, Lu J, Yang J, Shu Z, Shechtman E, and Huang J-B. Pose with style: Detail-preserving pose-guided image synthesis with conditional styleGAN. *ACM Trans Graph* 2021; 40(6): 1–11.
7. Wang TY, Ceylan D, Singh KK, and Mitra NJ. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In: *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 268–277.
8. Efros AA and Freeman WT. Image quilting for texture synthesis and transfer. In: *Seminal Graphics Papers: Pushing the Boundaries,* Vol. 2, 2023, pp. 571–576.
9. Wang Z, Zhao L, Chen H, et al. Texture reformer: Towards fast and universal interactive texture transfer. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2624–2632.
10. Wang B, Zheng H, Liang X, Chen Y, Lin L, and Yang M. Toward characteristic-preserving image-based virtual try-on network. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 589–604.
11. Zhou Z, Liu S, Han X, et al. Learning Flow Fields in Attention for Controllable Person Image Generation. *arXiv preprint arXiv:2412.08486* (2024).
12. Yan H, Zhang H, and Zhang Z. Learning to disentangle the colors, textures, and shapes of fashion items: A unified framework. *IEEE Trans Multimedia* 2023; 26: 5615–5629.
13. Shi J, Zhang H, Zhou D, and Zhang Z. Toward intelligent interactive design: A generation framework based on cross-domain fashion elements. In: *Proceedings of the 31st ACM International Conference on Multimedia*. New York: ACM Press, 2023, pp. 7152–7163.
14. Jung J, Kim H, and Park J. Deep fashion designer: Generative adversarial networks for fashion item generation based on many-to-one image translation. *Electronics* 2025; 14(2): 220.
15. Solanki A, Kang SS, Singla S, and Gururaja TS. High-resolution fashion image generation using Quantum-GAN. In: *2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*. IEEE, 2024, pp. 118–123.
16. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, vol. 27, 2014.
17. Sohl-Dickstein J, Weiss E, Maheswaranathan N, and Ganguli S. Deep unsupervised learning using

nonequilibrium thermodynamics. In: *International Conference on Machine Learning*, 2015, pp. 2256–2265.

18. Song Y and Ermon S. Generative modeling by estimating gradients of the data distribution. In: *Advances in Neural Information Processing Systems*, vol. 32, 2019.
19. Ho J, Jain A, and Abbeel P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.
20. Zhang Y, Zhang T, and Xie H. TexControl: Sketch-based two-stage fashion image generation using diffusion model. In: *2024 Nicograph International (NicoInt)*. IEEE, 2024, pp. 64–68.
21. Cao S, Chai W, Hao S, Zhang Y, Chen H, and Wang G. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Trans Multimedia* 2023; 26: 3962–3975.
22. Karagoz HF, Baykal G, Eksi IA, and Unal G. Textile pattern generation using diffusion models. *arXiv preprint arXiv:2304.00520* (2023).
23. Hu S, Zhang G, Zhao X, Li Z, and Li W. A method for yarn quality fluctuation prediction based on multi-correlation parameter feature subspace mechanism in spinning process. *J Eng Fibers Fabrics* 2023; 18: 15589250231208703.
24. Hu S and Zhang J. Modeling of fabric sewing break detection based on U-Net network. *Text Res J* 2024; 94(23–24): 2695–2706.
25. He W, Song B, Zhang N, Xiang J, and Pan R. Modeling and realization of image-based garment texture transfer. *Vis Comput* 2024; 40(9): 6063–6079.
26. Neverova N, Novotny D, Szafraniec M, Khalidov V, Labatut P, and Vedaldi A. Continuous surface embeddings. In: *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17258–17270.
27. Neverova N, Sanakoyeu A, Labatut P, Novotny D, and Vedaldi A. Discovering relationships between object categories via universal canonical maps. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 404–413.
28. Podell D, English Z, Lacey K, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
29. Kirillov A, Mintun E, Ravi N, et al. Segment anything. In :*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
30. Zhang J, Li H, Wang X, and Zhu Wei. Design of printed pattern generation method based on diffusion models. *Comput Meas Control* 2024; 32(10): 243–249, DOI: 10.16526/j.cnki.11-4762/tp.2024.10.035 (in Chinese).
31. Aigerman N and Groueix T. Generative Escher meshes. In: *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
32. Rodriguez-Pardo C and Garces E. SeamlessGAN: Self-supervised synthesis of tileable texture maps. *IEEE Trans Visualiz Comput Graph* 2022; 29(6): 2914–2925.
33. Ye H, Zhang J, Liu S, Han X, and Yang W. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
34. Zhang L, Rao A, and Agrawala M. Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
35. Aggarwal P. Fashion product images dataset. Kaggle, 2019. doi: 10.34740/KAGGLE/DS/139630.
36. Zhang R, Isola P, Efros AA, Shechtman E, and Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
37. Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 2004; 13(4): 600–612.
38. Heusel M, Ramsauer H, Unterthiner T, Nessler B, and Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Advances in Neural Information Processing Systems*, vol. 30, 2017.
39. ComfyUI: A Flexible User Interface for AI Image Generation. GitHub, https://github.com/comfyanonymous/ComfyUI (accessed 15 January 2025).
40. Face H. ControlNet-union-SDXL-1.0, https://huggingface.co/xinsir/controlnet-union-sdxl-1.0 (accessed 15 January 2025).
41. Kwon G and Ye JC. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264* (2022).
42. Xiang X, Liu D, Yang X, Zhu Y, Shen X, and Allebach JP. Adversarial open domain adaptation for sketch-to-photo synthesis. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,* 2022, pp. 1434–1444.
43. Kang H, Lee S, and Chui CK. Coherent line drawing. In *Proceedings of the 5th International Symposium on Non-photorealistic Animation and Rendering*, 2007, pp. 43–50.
44. Soria X, Li Y, Rouhani M, and Sappa AD. Tiny and efficient model for the edge detection generalization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1364–1373.
45. Canny J. A computational approach to edge detection. *IEEE Trans Pattern Anal Machine Intell* 1986; 6: 679–698.
46. Yang L, Kang B, Huang Z, et al. Depth Anything v2. In: *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 21875–21911.
47. Bhat SF, Birkl R, Wofk D, Wonka P, and Müller M. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).
48. Yang L, Kang B, Huang Z, Xu X, Feng J, and Zhao H. Depth Anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10371–10381.