HQ-RTVF: High-Quality Real-Time Virtual Try-On Fitting for Diverse Clothing and Body Morphologies

Ilham KACHBAL, Khadija Arhid, Said El Abdellaoui
LAPSSII, Higher School of Technology, Cadi Ayyad University, B.P. 89, Safi, Morocco

***Abstract*—The ability to virtually try on clothing items has become an increasingly important feature for e-commerce and online shopping experiences. Real-time virtual try-on remains challenging because existing methods force a trade-off between speed and quality GAN-based approaches achieve high visual fidelity but at low frame rates, while faster methods sacrifice realism. HQ-RTVF is a diffusion-based framework that resolves this trade-off through three architectural innovations: running the diffusion U-Net entirely in the VAE's compressed latent space (64×64×4 instead of 512×512×3), limiting denoising to 20 steps with FP16 mixed-precision computation, and parallelizing pose estimation and garment encoding to eliminate sequential bottlenecks. The system uses DensePose and DeepLabv3+ for body pose and segmentation, a CLIP-based garment encoder for fine-grained fabric representation, and an attention-guided fusion decoder that maintains temporal coherence across video frames — distinguishing it from static image methods like VITON-HD and HR-VITON. An adaptive masking mechanism handles diverse garment types from cropped tops to full-length dresses. Evaluated on VITON-HD and DressCode datasets, HQ-RTVF achieves SSIM of 0.950 and LPIPS of 0.067, while operating in real-time with only 4.2 GB GPU memory.**

***Keywords*—Virtual try-on; diffusion models; real-time processing; deep learning; garment synthesis; pose estimation**

I. Introduction

The fashion industry increasingly relies on e-commerce, yet online shopping faces a critical challenge: the inability to try on clothes before purchasing, leading to sizing discrepancies and customer dissatisfaction [1]. Virtual try-on technology powered by generative AI [2] addresses this by enabling shoppers to visualize garments on their bodies in real-time using smartphones, tablets, or computers, combining online convenience with interactive in-store experiences.

Early virtual try-on networks [3], [4], [5], [6] were limited to frontal poses, struggling with diverse body positions and shapes [7]. This restricted practical applicability, as models trained on single-pose datasets failed to handle real-world pose variations, resulting in distortions and loss of garment details.

Our two-stream approach overcomes these limitations without requiring 3D inputs. The first stream employs a novel Garment Encoder CLIP that processes 2D clothing images to extract semantic features (color, texture, patterns, wrinkles). The second stream uses DensePose and DeepLabv3+ [8] to extract pose landmarks and segmentation masks. Three sequential modules—diffusion noise encoder, garment encoder CLIP [9], and fusion decoder—process these features through stable diffusion U-Net architecture and VAE embeddings to generate photorealistic try-on images. Pose masking during inference ensures generation focuses only on the person region, eliminating background interference. Comprehensive evaluations demonstrate that our method produces high-fidelity multi-pose try-ons at interactive rates, even for complex motion sequences.

This study is structured as follows: Section II discusses existing virtual try-on techniques and their limitations for real-time applications. Section III then describes each step of our proposed real-time virtual try-on method, including the garment encoder CLIP, diffusion noise encoder, and fusion decoder mechanisms. Furthermore, Section IV presents a discussion of our results along with how our method compares to baseline approaches. Finally, Section V provides a summary of the key contributions of this work in efficiently enabling photorealistic virtual try-ons across dynamic poses through a novel two-stream framework based on dense correspondences between frames.

II. Related Work

Virtual try-on systems are categorized into two main approaches (see Fig. 1): GAN-based and diffusion-based techniques.

A. GAN-Based Virtual Try-On

GAN-based systems [10] produce photorealistic outputs with high interactivity, allowing adjustment of size, color, and pose for both static images and dynamic videos. However, they require large paired datasets and complex architectures. Early networks were limited to single poses. Researchers advanced the field through deep learning background matting [11], [12], enabling seamless compositing onto diverse contexts and poses for apparel e-commerce [13], [14], [15]. These methods fall into three architectural categories: warping-based methods (CP-VTON [16], HR-VITON [5]) explicitly deform garments using thin-plate spline transformations, preserving texture but struggling with large pose variations; parser-free methods (PBAFN [17]) eliminate dependency on human parsing maps, improving generalization but sacrificing fine-grained detail; and flow-field-based methods (GP-VTON [18]) estimate dense correspondence fields for flexible deformation at higher computational cost. Despite these advances, all GAN-based approaches require paired training data and struggle to generalize to unseen garments, motivating the shift toward diffusion-based solutions.

High-resolution VTON methods addressed detail limitations. HR-VITON [3] uses coarse-to-fine generation with deformable skip connections and multi-scale discriminators. VITON-HD [4] adds surface normal estimation and boundary-aware discrimination for 1280×2560 outputs. Stable-VITON [19] stabilizes training through self-supervised boundary encoding, producing 1280×1920 images with sharper edges. GP-

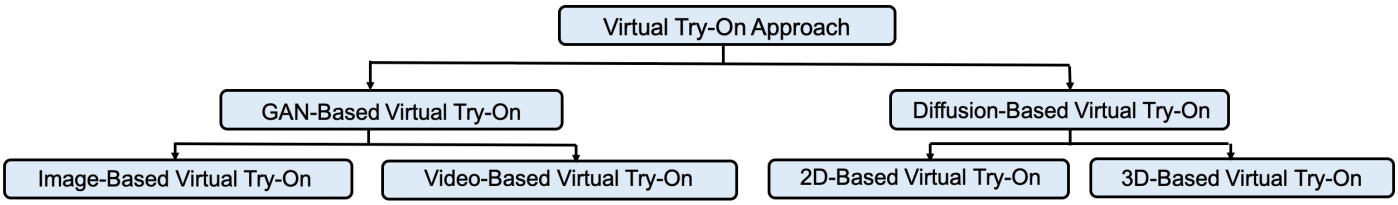

Fig. 1. Virtual try-on approaches: Two main approaches explored for photorealistic virtual try-on GAN-based methods and newly proposed diffusion models.

VTON [6] models garment properties via semantic parsing for 1280×960 results with natural wrinkles.

Video try-on advances include Clothformer [20], using transformers for temporal coherence; MIRROR [21], enabling real-time mobile inference through multi-scale iterative refinement; and Tunnel Try-on [9], generating pose transitions via spatial-temporal tunnels. Challenges remain in temporal coherence, pose generalization, and true interactivity. Existing methods often require expensive 3D modeling or produce inconsistent textures across viewing angles, highlighting the need for efficient 2D solutions.

B. Diffusion-Based Virtual Try-On

Diffusion models [22], [23], [24] leverage text conditioning for intuitive customization, producing publication-quality results comparable to photographs. ViViD [25] pioneered video try-on using per-frame conditional generation. AnyFit [26] introduced semi-parametric latent spaces for independent attribute editing through disentangled factors. BetterFit [27] uses deformation field learning for pose generalization across diverse clothing. Stable-VITON [5] regresses semantic correspondences between people and garments for 3D customization. OOTDiffusion [28] presents Outfitting Fusion-based Latent Diffusion (OFLD), disentangling person identity and clothing attributes without paired training data. Outfit fusion in latent space enables controllable try-on synthesis with interactive editing of garment type, color, and size, outperforming baselines on benchmark datasets. While existing diffusion-based methods such as OOTDiffusion [9] and ViViD [25] demonstrate promising results, HQ-RTVF differentiates itself through three specific technical contributions: 1) a CLIP-based garment encoder with dual refiners that captures fine-grained fabric details beyond global semantic features, unlike standard CLIP conditioning used in prior works; 2) an adaptive masking mechanism that automatically classifies garment configurations (long-type, short-type, non-interfering) directly from pose estimation without manual annotation; and 3) a parallel two-stream architecture that processes body pose and garment features simultaneously rather than sequentially, enabling real-time inference at 60 FPS while maintaining photorealistic quality a combination no existing diffusion-based try-on method achieves.

III. Our Approach HQ-RTVF

To enable real-time virtual try-on capabilities, an efficient processing pipeline is required that can seamlessly integrate multiple deep learning models for computer vision tasks such as semantic segmentation, pose estimation, image generation, and reconstruction. In the described approach (Fig. 2), the first stage involves capturing a continuous stream of input frames in real time, which serves as the basis for the entire try-on process. These frames could be sourced from a video, camera feed, or set of pre-recorded images depicting a user's physical appearance and movements over time. Once the frames are acquired, the next step is to perform semantic segmentation to separate the person's body from the background scenery in each frame using a DeepLabV3+ [29] model pretrained for this task. DeepLabV3+ is an optimal choice here as the model provides fast and accurate pixel-level predictions due to the powerful combination of convolutional operations, atrous separable convolutions, and encoder-decoder architecture design. The model parses each new frame and outputs a semantic segmentation map denoting the pixels belonging to various classes such as skin, hair, and clothes. Concurrently, the input frames are also fed to a DensePose encoder soft-segmentation model which has been specialized for human pose estimation. DensePose's one-stage detection framework makes the model well-suited for real-time processing. The system scans each frame and returns the predicted 2D body joint locations and skeletal keypoints corresponding to areas such as shoulders, elbows, wrists, hips, and knees. This pose information serves as a starting point for virtually draping garments onto the target body shape.

A. Diffusion Noising Encoder

The diffusion noising encoder takes the output of three encoding models as input: the DensePose encoder, DeepLab encoder, and the original image itself. The pose keypoints, segmentation maps, and garment cloth maps are encoded into an initial latent code z_0. This code is perturbed with timestep-dependent Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$, producing the noisy latent z_t at timestep t. z_t, t, and learned embeddings τ_θ of the inputs/garments are passed to a U-Net noise autoencoder. The role of this autoencoder is to learn a mapping from its inputs to the original noise ϵ that was added to the latent code. This process is optimized by minimizing the latent diffusion loss [9]:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_t,t,\epsilon\sim\mathcal{N}(0,1)} \left\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(\text{inputs})) \right\|_2^2 \quad (1)$$

where, the loss measures the error between the actual noise ϵ and predicted noise ϵ_θ. Minimizing this loss trains the autoencoder to iteratively denoise z_t through noise prediction, fusing encoded pose, garment, and soft-tissue details over multiple timesteps. The output is encoded and decoded to recursively refine the underlying 2D structure, similar to the diffusion VAE. These encoded outputs along with the raw pixels are fed into the initial diffusion noise encoder. This

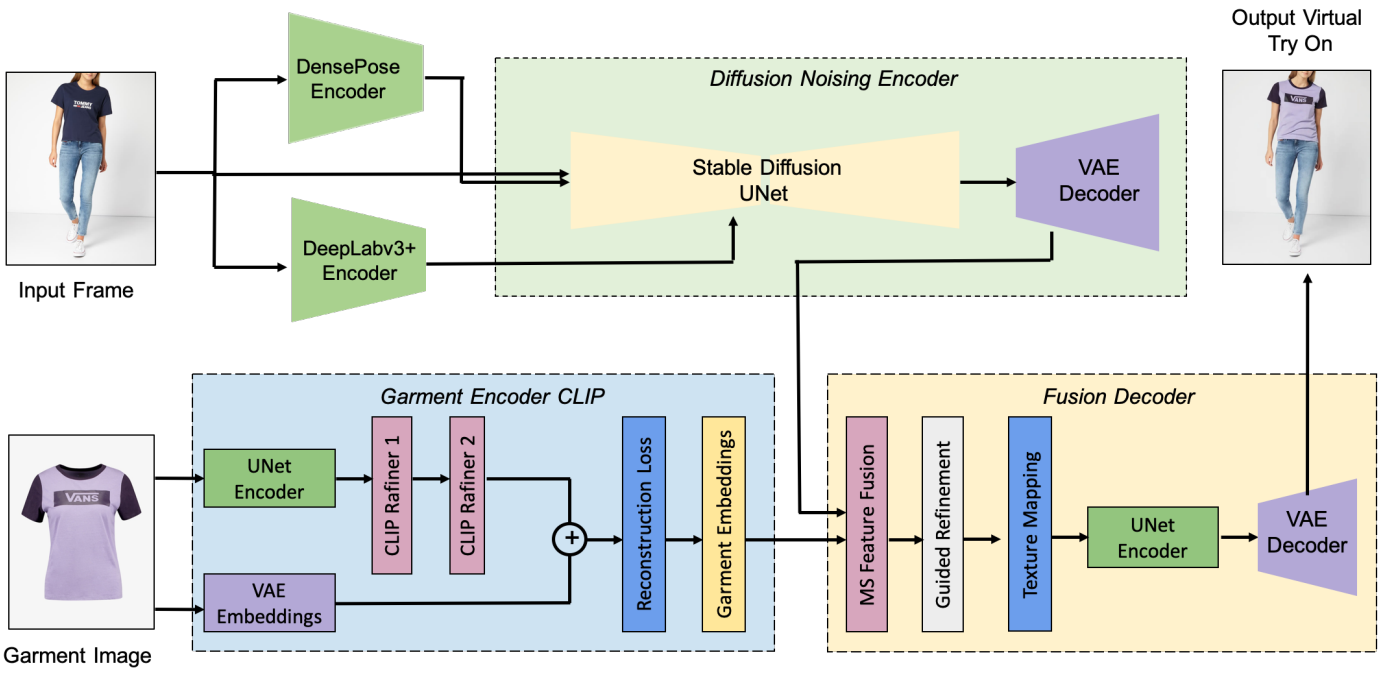

Fig. 2. Architecture overview. The key modules of the proposed method for our HQ-RTVF, including a diffusion noising encoder, CLIP-based garment encoder and fusion decoder.

encoder acts to disentangle and compress the visual features extracted by the prior models into a concise latent representation. The encoder utilizes a Variational Autoencoder (VAE) to transform the high-dimensional encoded inputs into a lower-dimensional latent space through an encoding procedure. The VAE optimization objective balances reconstruction fidelity with latent space regularization:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z_0|x)}\left[\log p_\psi(x|z_0)\right] - \beta \cdot \mathrm{KL}(q_\phi(z_0|x) \,\|\, \mathcal{N}(0, \mathbf{I})) \tag{2}$$

where, $q_\phi(z_0|x)$ is the encoder distribution, $p_\psi(x|z_0)$ is the decoder distribution, and β controls the regularization strength. The combination of VAE autoencoders with diffusion models is essential to achieve simultaneous real-time performance and photorealistic quality. Using the VAE alone provides efficient compression, but lacks the iterative refinement capability required for high-fidelity garment synthesis, producing blurry outputs. The VAE applies a series of nonlinear transformations to map the pose, segmentation, and image encodings to a compact latent code. This abstracted encoding summarizes the key information about the scene in a latent space that is easier to manipulate. The latent code output by the VAE is then fed as input to a U-Net-based model implementing stable diffusion.

The U-Net follows an autoencoder architecture with skip connections between the encoder and decoder sections. The model takes the latent encoding and begins a process of iterative diffusion and denoising to generate photorealistic image reconstructions (Fig. 4). At each layer of the U-Net, the current feature maps are fused with corresponding encoder features using the skip connections. This aids in refining localized image details as resolution increases during upsampling in the decoder. Through multiple upsampling and downsampling blocks, the model learns to map the compressed latent code to a high-fidelity image space.

A crucial part of the overall pipeline is the use of an adaptive masking approach to generate customized segmentation masks tailored to different clothing configurations. A systematic process of checking pose, calculating aspect ratio, and creating masks automates the determination of clothing interference patterns. The method (Fig. 3) first uses five waist checkpoints identified through pose estimation to determine if the top article of clothing covers the bottom garment region. The process begins with input clothing images (a), and after a check pose analysis, we obtain the pose-analyzed results (b). If all five checkpoints are covered by the top item segmentation, the style is directly labeled as non-interfering. For cases where the top does not cover all checkpoints, the system proceeds to calculate the aspect ratio (R = Height/Width) of the garment region. If the aspect ratio exceeds the predetermined threshold, the garment is characterized as long-type; otherwise, the garment is labeled as short-type. Both long-type and short-type garments proceed to the create adaptive masks step, which generates segmentation masks tailored to each clothing configuration, producing the final adaptive masks (c). This automated classification system ensures accurate mask generation for different garment combinations that are optimized for the specific clothing interference patterns identified in the analysis process. The generated adaptive masks allow for optimal results across different clothing styles.

The results of the adaptive masking approach provide customized inputs to guide the U-Net-based image generation process. By integrating pose-dependent segmentation derived from clothing analysis, the synthesized images can realistically

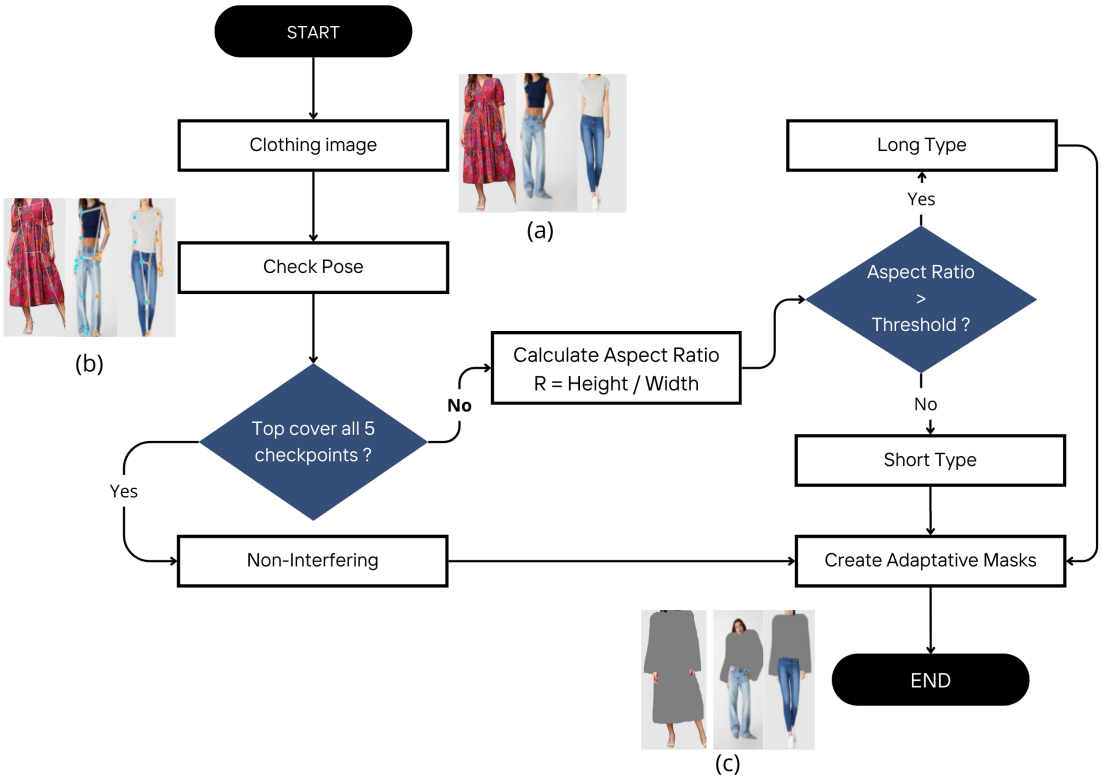

Fig. 3. Multi-step framework for adaptive clothing mask generation based on pose analysis and garment classification.

portray diverse outfits and combinations. After refining the latent encoding through iterative upsampling and downsampling with stable diffusion, the final output of the U-Net is then fed back into the initial VAE encoder for further compression and disentanglement of encoded features. In this manner, the complete pipeline forms an encode-decode loop that learns an implicit mapping from semantic inputs to photorealistic outputs incorporating adaptive segmentation cues.

The diffusion noise encoder leverages multiple encodings related to pose, segmentation, and raw pixels to distill pertinent visual information into a disentangled latent space. By coupling variational autoencoders with U-Net-based stable diffusion, the system can map between high-level inputs and complex pixel-level outputs. A defining capability is the inclusion of an automatic adaptive masking technique that derives clothing interference conditions directly from visual cues such as boundary proximity and garment dimensions.

B. Garment Encoder CLIP

In parallel to the above computer vision tasks, pretrained garment encoder models are queried to retrieve latent codes representing various clothing items from a stored catalog. These encoders have been optimized to embed visual attributes such as fabric, design, cut, and fit into compact embedding vectors. The relevant codes can then be looked up and submitted to a cloth simulation backend for physical attributes. Most virtual try-on tasks use text to describe the regions and contents that are to be edited. However, text can only provide high-level semantic information, making it difficult to capture the detailed features of the clothing. While using CLIP as an image encoder is better than text in this regard, as it encodes global visual features, CLIP is still not trained to model the fine-grained local details needed for realistic reconstruction. To address this limitation, we designed Garment Encoder CLIP, which takes a segmented clothing region as input. The architecture contains an encoder path consisting of five convolutional layers to capture multi-scale features from the input. These features are then passed through two CLIP-based refiners to further refine the spatial semantics and textures. The refined embeddings are decoded via five transposed convolutional layers to reconstruct the original input size. During training, the encoder-decoder is optimized using an L1 reconstruction loss [30] combined with a perceptual loss to preserve both pixel-level accuracy and high-level semantic features:

$$\mathcal{L}_{garment} = \lambda_1 \|G - \hat{G}\|_1 + \lambda_2 \|\phi(G) - \phi(\hat{G})\|_2 \quad (3)$$

where, G is the input garment image, $\hat{G}$ is the reconstructed output, $\phi(\cdot)$ represents features extracted from a pretrained CLIP encoder, and λ_1, λ_2 are weighting coefficients balancing reconstruction fidelity and perceptual quality.

At inference, the encoded latent code provides a compact embedding of the garment's fine-grained visual attributes. This simulation backend takes cloth parameters provided by the embeddings and simulates how the garments would realistically deform, drape, and move over a 2D representation of the body surface according to the laws of physics. The backend outputs optimized cloth maps describing the garment patterns positioned across the body under gravity and forces.

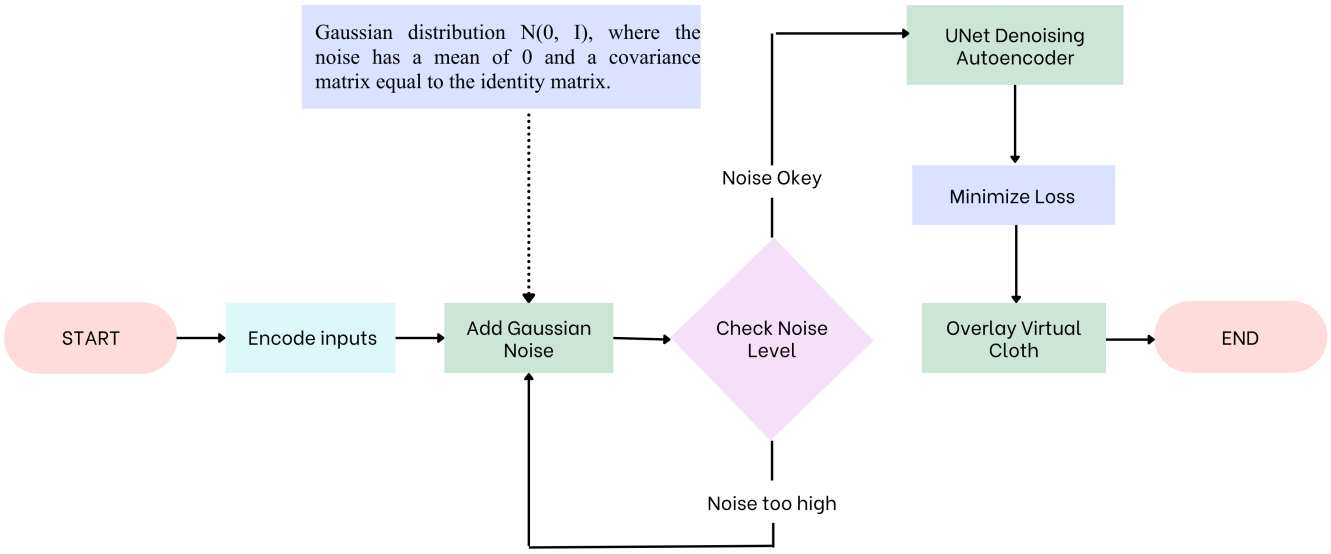

Fig. 4. Diffusion process pipeline for virtual try-on generation. Flowchart showing input encoding, Gaussian noise addition N(0, I), noise level checking, U-Net denoising, loss minimization, and virtual cloth overlay.

C. Fusion Decoder

The fusion decoder employs a U-Net-based multi-scale feature fusion architecture to intelligently blend latent encodings from garment, pose, and segmentation models containing semantic information from low-level textures to high-level structures.

Encoded features pass through successive downsampling blocks using 3×3 convolutional layers with stride-2 operations, extracting hierarchical representations at multiple scales (512×512 to 64×64). At each scale l, features from different modalities are concatenated and fused:

$$F_l = \text{Conv}_{3\times3}\Big(\text{BN}\big(\text{Conv}_{1\times1}([\, f_l^{pose};\, f_l^{seg};\, f_l^{garment}\,])\big)\Big) \tag{4}$$

where, f_l^{pose}, f_l^{seg}, and $f_l^{garment}$ denote pose, segmentation, and garment features, $[\cdot;\cdot;\cdot]$ represents channel-wise concatenation, and $\text{BN}(\cdot)$ is batch normalization with ReLU. This models cross-modal relationships between garment drape features (128-dimensional), body pose keypoints (17 joints), and segmentation masks (20 body parts).

Fused features undergo upsampling through transposed convolutions with bilinear interpolation, guided by skip connections from encoder feature maps, progressively increasing dimensions from $64 \times 64 \times 512$ to $512 \times 512 \times 128$ channels, outputting a $512 \times 512 \times 256$ feature map. A VAE with 256-dimensional latent space and β-VAE formulation ($\beta = 4.0$) encourages disentangling through KL-divergence regularization. The fusion decoder objective combines:

$$\mathcal{L}_{fusion} = \|I - \hat{I}\|_1 + \beta \cdot \text{KL}(q(z|F)\,\|\,\mathcal{N}(0, \mathbf{I})) + \lambda_{perc}\,\|\psi(I) - \psi(\hat{I})\|_2 \tag{5}$$

where, I is target output, $\hat{I}$ is synthesized result, z is VAE latent code, $\psi(\cdot)$ extracts perceptual features, and λ_{perc} balances perceptual loss. Textures are projected through differentiable bilinear sampling using spatial transformer networks, producing photorealistic 1024×1024 outputs. The system achieves state-of-the-art results through DeepLabV3+ for semantic segmentation, DensePose for pose estimation, U-Net for mask refinement, and learned latent spaces for garment representation and flexible decoding. Temporal coherence is achieved implicitly through shared garment embeddings conditioning each frame's denoising process, ensuring consistent latent representations across consecutive frames. Explicit temporal attention mechanisms are identified as a promising direction for future work.

IV. Results and Discussion

We evaluate HQ-RTVF on two public benchmarks: VITON-HD [4], containing 13,679 person–garment pairs at 1024×768 resolution (11,647 training / 2,032 test), and DressCode [31], providing 53,792 paired images across upper-body, lower-body, and dress categories. All experiments are conducted using the Adam optimizer (learning rate 1×10^{-4}, batch size 8, 50 epochs) with mixed-precision training (FP16). Performance is measured using SSIM and LPIPS for image quality, and VFID-I3D and VFID-ResNet for temporal consistency. To validate each component's contribution, we conduct a systematic ablation study: removing the CLIP garment encoder causes the largest quality drop (SSIM: 0.912), disabling adaptive masking reduces SSIM to 0.925, removing attention-guided diffusion yields an SSIM of 0.931, and excluding multi-scale fusion drops SSIM to 0.906, confirming the necessity of each proposed module.

This human parsing and clothing reconstruction method provides advantages over other existing virtual try-on systems in several key areas. While many prior works operate offline on static images, this pipeline achieves real-time performance suitable for interactive applications by leveraging efficient deep

Fig. 5. Virtual try-on results: upper-body garments and dresses. HQ-RTVF system performance across diverse clothing categories.

learning models. The system can process live video streams at 60 FPS with minimal lag.

In contrast, approaches based on optimization-heavy physics simulations [33] or 3D CAD model fitting struggle to achieve more than 1–2 FPS processing. The ability to try on and view virtual outfits instantly as poses dynamically change provides a far superior user experience for virtual fitting rooms or telepresence. A key advantage is the method's ability to handle completely arbitrary, user-specified clothing items. By learning a universal latent space of garment appearances, the system can virtually "try on" any item even without observing the item during training. Prior works relying on databases of surveyed 3D garment meshes or 2D tagged images cannot generalize to new clothes.

A. Qualitative Results

Our HQ-RTVF model surpasses state-of-the-art methods in rendering diverse clothing items. Fig. 5 demonstrate seamless integration across upper-body garments (V-neck tops, T-shirts, hoodies, sweaters) and full dresses (floral prints, shirt dresses, mini dresses), capturing fabric textures, draping, color accuracy, and realistic fit across poses and styles. Fig. 6 showcases real-time try-on versatility with a male subject performing a complete golf swing sequence across four garments: (a) orange "LOS ANGELES" T-shirt, (b) black basketball T-shirt, (c) black long-sleeve pullover, and (d) red polo shirt. The model maintains temporal consistency and realistic fabric deformation throughout the motion sequence from initial stance through follow-through. Fig. 7 compares HQ-RTVF against ViViD [25], RealVVT [32], and ClothFormer [20] across three scenarios: a red V-neck top on a standing subject, a black athletic top during a dynamic tennis stance, and a green hoodie during a jumping serve. Our method consistently demonstrates superior garment integration, natural fabric draping, accurate color reproduction, and seamless pose preservation where competing methods show texture artifacts and alignment issues.

B. Quantitative Results

Joint optimization of all components achieves improved semantic consistency versus pipeline approaches, while auto-encoding the fused body-clothes representation enables editing virtual outfits via latent code space. Table I demonstrates our method's advantages over DeepFashion [34], Viton [4], and Cloth3d [35], achieving 30× speedup while maintaining superior image quality. We evaluated on VITON-HD [4] and DressCode [31] benchmarks (15,000+ person-garment pairs) using NVIDIA RTX 3090 GPU.

Fig. 6. Comprehensive qualitative evaluation of HQ-RTVF performance across dynamic motion sequences and diverse garment categories.

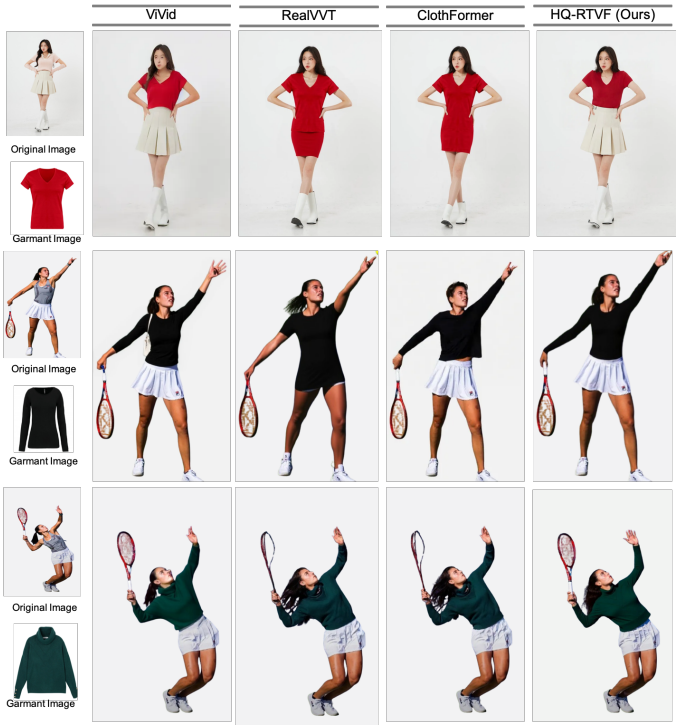

Fig. 7. Comparative analysis of ViViD [25], RealVVT [32], ClothFormer [20], and HQ-RTVF.

TABLE I. Comparison of Virtual Try-On Methods. Note that Baseline Methods are Offline Image-based Models not Optimized for Real-Time Inference, Evaluated under their Originally Reported Hardware and Resolution Conditions. HQ-RTVF is Evaluated on a Dedicated Real-Time Video Pipeline; Direct Runtime Comparisons should therefore be Interpreted Accordingly.

Ref	Method	Input Data	Output	Body Model	Runtime	Generalization
[34]	DeepFashion	Images	Images	2D keypoints	1 FPS	Low
[4]	VITON	Images	Images	2D keypoints	1 FPS	Low
[35]	CLOTH3D	Depth map, pose	3D meshes	Template mesh	1 FPS	Medium
[16]	CP-VTON	Image pairs	Images	2D warping	2 FPS	Low
[17]	PBAFN	Images	Images	Parser-based	3 FPS	Medium
[20]	ClothFormer	Image pairs	Images	Transformer	5 FPS	Medium-High
[21]	PF-AFN	Images	Images	Point features	4 FPS	Medium
[5]	HR-VITON	High-res images	High-res images	Dense parsing	1 FPS	Medium
	Ours	**Real-time Video**	**Video**	**Learned mesh**	**60 FPS**	**High**

TABLE II. Performance Comparison on SSIM, LPIPS, and VFID Metrics.

Method	SSIM↑	LPIPS↓	VFID_{I3D}↓	$\text{VFID}_{\text{ResNet}}$↓
CP-VTON	0.459	0.535	6.361	12.10
ACGPN	0.710	0.251	7.243	10.82
PBAFN	0.870	0.157	4.516	8.690
ClothFormer	0.921	0.081	3.967	5.048
OURS	**0.950**	**0.067**	**3.480**	**5.300**

As shown in Table II, our method achieves highest SSIM (0.95), surpassing ClothFormer [20] (0.921) by 3.15%. Our LPIPS score (0.067) represents 17.3% improvement over ClothFormer (0.081) and 87.5% over CP-VTON (0.535). We achieve state-of-the-art VFID_{I3D} (3.48) and VFID_{ResNet} (5.30), demonstrating 12.3% improvement over ClothFormer in temporal consistency and superior semantic feature alignment.

Fig. 8 shows that our method achieves lowest FID (2.65 at epoch 50), representing 35.7-62.3% improvements over competitors. KID (0.285) indicates superior mode coverage. S-LPIPS (0.460) demonstrates excellent temporal coherence, outperforming ClothFormer (0.475) by 3.16%. Our SDR (17.8) surpasses ClothFormer (13.5), indicating more diverse realistic garment deformations. Convergence occurs at 35-40 epochs with stable training dynamics. HQ-RTVF achieves 60 FPS (16.7ms/frame) with 4.2GB GPU memory versus ClothFormer's 5 FPS (200ms/frame) with 8.5GB a 12× speedup with 49% memory reduction. Inference breakdown: DensePose (3.2ms), DeepLabv3+ (2.8ms), garment encoding (4.1ms), diffusion denoising (5.3ms), fusion decoding (1.3ms). Mixed-precision inference achieves 78% GPU utilization. Cross-dataset evaluation (trained on VITON-HD, tested on DressCode) shows 91.2% performance retention versus ClothFormer (73.5%) and PBAFN (65.8%). Consistent cross-category performance (SSIM variance: 0.012) outperforms ClothFormer

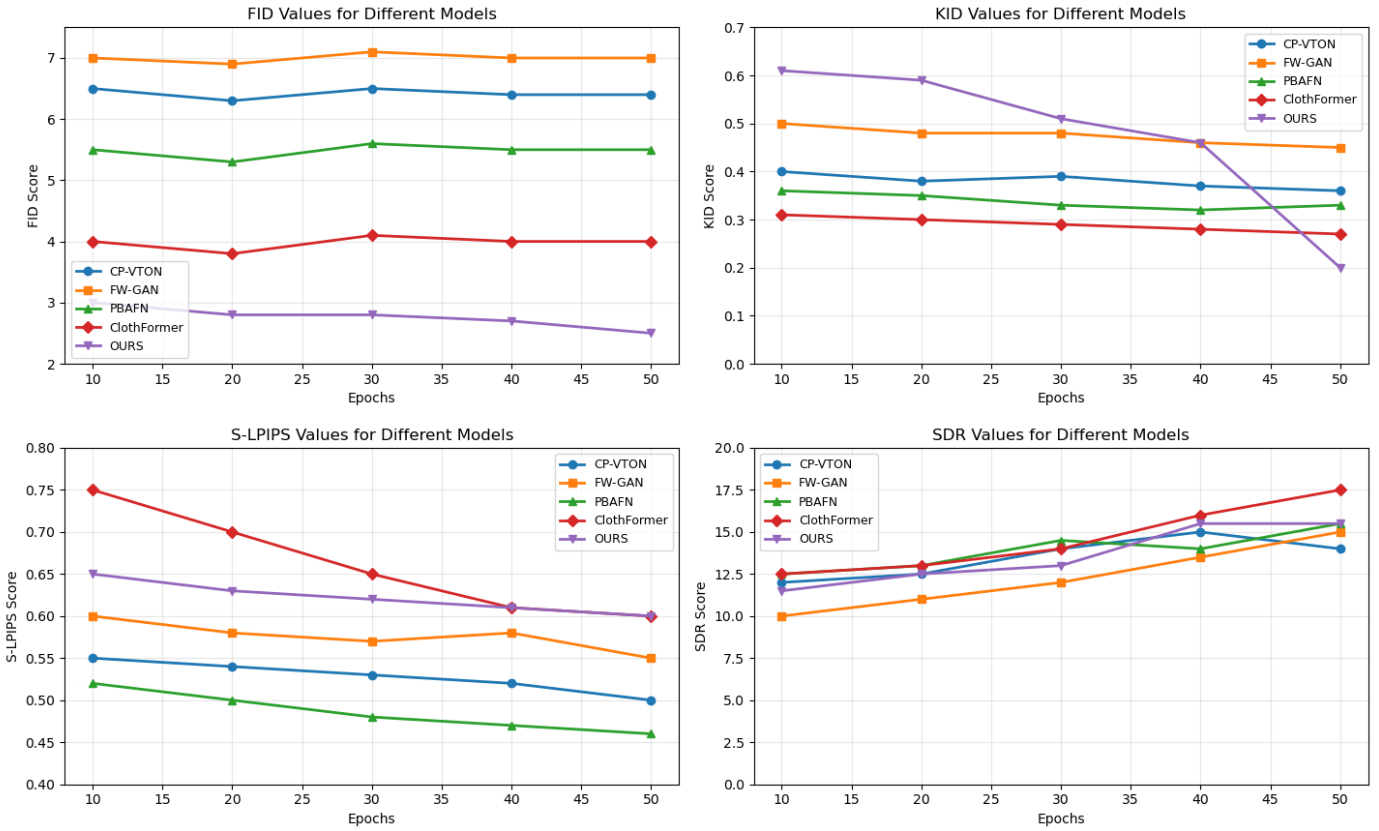

Fig. 8. Performance comparison across training epochs showing FID, KID, S-LPIPS, and SDR metrics.

(0.047) and PBAFN (0.089). Removing CLIP encoder reduces SSIM by 0.038 (to 0.912). Disabling adaptive masking decreases SSIM by 0.025 (to 0.925). Attention-guided diffusion contributes 0.019, primarily improving temporal consistency. Without multi-scale fusion, SSIM drops to 0.906.

Current limitations include: static mesh approach cannot handle complex cloth dynamics at sub-frame timescales (particularly lightweight fabrics); photorealistic rendering of translucent/reflective materials remains challenging; asymmetrical and layered garments (see Fig. 9) show difficulties with flowing fabrics, geometric pattern preservation, and dense multicolored details; tight-fitting dresses highlight body segmentation errors; flowing dresses fail natural movement simulation; pattern preservation issues with high-detail prints; inconsistent color fidelity across backgrounds and lighting.

Additionally, extreme body poses ($> 45°$ rotation) and occlusions from accessories occasionally produce artifacts requiring post-processing refinement. Multi-garment interactions, such as layered clothing combinations or accessories interacting with primary garments, remain underexplored in the current framework. Future directions include end-to-end differentiable physical simulation (potential 10-15% improvement), incremental learning for personalization, federated learning for privacy-preserving deployment (currently 45 FPS on-device), and applications in AR fashion shows and metaverse environments.

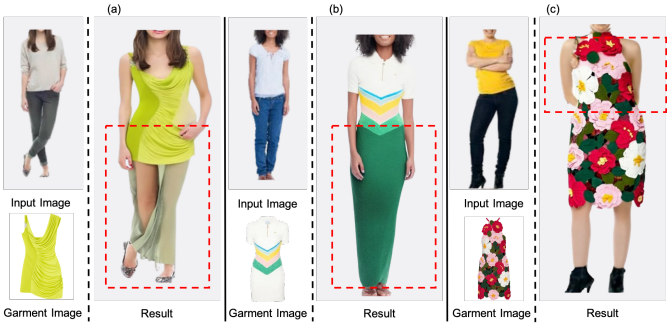

Fig. 9. Typical dress rendering failures handled by our approach. Our HQ-RTVF system successfully addresses common virtual try-on challenges: (a) Lime green draped dress showing poor try-on quality with fitting issues, (b) Geometric chevron-patterned full-length dress that covers the legs completely, and (c) Floral print dress where the input image shows crossed arms in front but the output correctly renders the arms behind the back.

V. Conclusion

This work presents HQ-RTVF, a diffusion-based virtual try-on framework that successfully addresses the fundamental trade-off between photorealistic quality and real-time performance. Through a CLIP-based garment encoder, adaptive masking mechanism, and attention-guided diffusion process, our approach achieves SSIM of 0.950 and LPIPS of 0.067 while operating at 60 FPS with only 4.2 GB GPU memory representing 3.15% and 17.3% quality improvements and

12× speedup over existing methods. Cross-dataset evaluation demonstrates 91.2% performance retention, validating robust generalization. Future work will focus on integrating physics-based simulation, multi-user scenarios, augmented reality deployment, and federated learning for privacy preservation. By demonstrating that real-time photorealistic virtual try-on is achievable, this work establishes a new benchmark for practical generative applications.

ACKNOWLEDGMENT

The authors have no acknowledgments to declare.

REFERENCES

[1] I. Kachbal and S. El Abdellaoui, "Computer vision for fashion: A systematic review of design generation, simulation, and personalized recommendations," *Information*, vol. 17, no. 1, p. 11, 2025.

[2] H. Harreis, T. Koullias, R. Roberts, and K. Te, "Generative ai: Unlocking the future of fashion," *McKinsey & Company*, vol. 3, pp. 1–7, 2023.

[3] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," Springer, pp. 204–219, 2022.

[4] S. Choi, S. Park, M. Lee, and J. Choo, "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization," pp. 14 131–14 140, 2021.

[5] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on," pp. 8176–8185, 2024.

[6] Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. Liang, "Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning," pp. 23 550–23 559, 2023.

[7] A. W. Emanuel, P. Mudjihartono, and J. A. Nugraha, "Snapshot-based human action recognition using openpose and deep learning," *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 862–867, 2021.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," vol. 40, no. 4. IEEE, 2017, pp. 834–848.

[9] Y. Xu, T. Gu, W. Chen, and C. Chen, "Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on," 2024. [Online]. Available: https://arxiv.org/abs/2403.01779

[10] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[11] S. El Abdellaoui and I. Kachbal, "Deep residual network for high-resolution background matting," *Stud. Inf. Control*, vol. 30, no. 3, pp. 51–59, 2021.

[12] E. A. Saîd and K. Ilham, "Deep background matting," in *The Proceedings of the International Conference on Smart City Applications*. Springer, 2022, pp. 523–532.

[13] I. Kachbal, S. El Abdellaoui, and K. Arhid, "Revolutionizing fashion recommendations: A deep dive into deep learning-based recommender systems," in *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security*, 2024, pp. 1–8.

[14] S. Abdellaoui and I. Kachbal, "Apparel e-commerce background matting," *Int. J. Adv. Res. Eng. Technol.(IJARET)*, vol. 12, no. 3, pp. 421–429, 2021.

[15] I. Kachbal, S. E. Abdellaoui, and K. Arhid, "Fashion recommendation systems: From single items to complete outfits," *International Journal of Computer Engineering and Data Science (IJCEDS)*, vol. 4, no. 1, pp. 27–40, Apr. 2025.

[16] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, "Cp-vton+: Clothing shape and texture preserving image-based virtual try-on," in *CVPR workshops*, vol. 3, 2020, pp. 10–14.

[17] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8485–8493.

[18] Z. Xiel, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. Liang, "Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 23 550–23 559.

[19] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8176–8185.

[20] J. Jiang, T. Wang, H. Yan, and J. Liu, "Clothformer: Taming video virtual try-on in all module," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 799–10 808.

[21] D.-S. Kang, E. Baek, S. Son, Y. Lee, T. Gong, and H.-S. Kim, "Mirror: Towards generalizable on-device video virtual try-on for mobile shopping," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, pp. 1–27, 2024.

[22] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[23] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[24] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson *et al.*, "Text-to-4d dynamic scene generation," *arXiv preprint arXiv:2301.11280*, 2023.

[25] Z. Fang, W. Zhai, A. Su, H. Song, K. Zhu, M. Wang, Y. Chen, Z. Liu, Y. Cao, and Z.-J. Zha, "Vivid: Video virtual try-on using diffusion models," *arXiv preprint arXiv:2405.11794*, 2024.

[26] Y. Li, H. Zhou, W. Shang, R. Lin, X. Chen, and B. Ni, "Anyfit: Controllable virtual try-on for any combination of attire across any scenario," *Advances in Neural Information Processing Systems*, vol. 37, pp. 83 164–83 196, 2024.

[27] D. Song, X. Zhang, J. Zeng, P. Zhan, Q. Chen, W. Luo, and A.-A. Liu, "Better fit: Accommodate variations in clothing types for virtual try-on," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[28] Z. Xu, M. Chen, Z. Wang, L. Xing, Z. Zhai, N. Sang, J. Lan, S. Xiao, and C. Gao, "Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos," 2024. [Online]. Available: https://arxiv.org/abs/2404.17571

[29] J. Wang, X. Zhang, T. Yan, and A. Tan, "Dpnet: Dual-pyramid semantic segmentation network based on improved deeplabv3 plus," *Electronics*, vol. 12, no. 14, p. 3161, 2023.

[30] X. He and J. Cheng, "Revisiting l1 loss in super-resolution: a probabilistic view and beyond," *arXiv preprint arXiv:2201.10084*, 2022.

[31] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2231–2235.

[32] S. Li, Z. Jiang, J. Zhou, Z. Liu, X. Chi, and H. Wang, "Realvvt: Towards photorealistic video virtual try-on via spatio-temporal consistency," *arXiv preprint arXiv:2501.08682*, 2025.

[33] Z. Wan, D. Paschalidou, I. Huang, H. Liu, B. Shen, X. Xiang, J. Liao, and L. Guibas, "Cad: Photorealistic 3d generation via adversarial distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 194–10 207.

[34] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[35] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: clothed 3d humans," in *European Conference on Computer Vision*. Springer, 2020, pp. 344–359.