EfficientVITON: An Efficient Virtual Try-On Model using Optimized Diffusion Process

Mostafa Kotb*, Mariam Ayman*, Ahmed Rashed*, Ashrakat Saeed*, Abdelrahman Said*, Ahmed Fares*
*Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt.

***Abstract*—Wouldn't it be much more convenient for everyone to try on clothes by only looking into a mirror? The answer to that problem is "virtual try-on," enabling users to digitally experiment with outfits. The core challenge lies in realistic image-to-image translation, where clothing must fit diverse human forms, poses, and figures. Early methods used 2D transformations in order to offer speed, but image quality was often disappointing and lacked the nuance of deep learning. Though GAN-based techniques enhanced realism, their dependence on paired data proved limiting. More adaptable methods offered great visuals but demanded significant computing power and time. Recent advances in diffusion models have shown promise for high-fidelity translation, yet the current crop of virtual try-on tools still struggle with detail loss and warping issues. To tackle these challenges, this thesis proposes EfficientVITON; a new virtual try-on system leveraging the impressive pretrained Stable Diffusion model for better images and deployment feasibility. The system includes a spatial encoder to maintain clothing's finer details and zero cross-attention blocks to capture the subtleties of how clothes fit a human body. Input images are carefully prepared, and the diffusion process has been tweaked to significantly cut generation time without image quality loss. The training process involves two distinct stages of fine-tuning, carefully incorporating a balance of loss functions to ensure both accurate try-on results and high-quality visuals. Rigorous testing on the VITON-HD dataset, supplemented with real-world examples, has demonstrated that EfficientVITON achieves state-of-the-art results. Ultimately, this thesis successfully demonstrates a method for leveraging pretrained diffusion models in virtual try-on, obtaining high quality results, while maintaining practicality and deployability within the fashion sector.**

***Index Terms*—Virtual Try-On, Diffusion Model, Stable Diffusion, Spatial Encoder, Zero Cross-Attention, EfficientVITON.**

1 Introduction

1.1 Motivation

The rapid growth of artificial intelligence (AI) and computer vision has brought significant changes to various industries, with fashion and retail being some of the most affected. Among the most innovative applications is virtual try-on technology, which enables users to see how garments would look on their own images. [1][2][3] By seamlessly integrating digital and physical shopping experiences, this technology has transformed e-commerce, providing customers with personalized, interactive, and efficient ways to shop.

Beyond improving customer satisfaction, virtual try-on systems tackle several key challenges in the fashion industry:

1) High Return Rates
   E-commerce platforms frequently face high return rates due to issues with sizing and fit; virtual try-on tools offer realistic previews, helping customers make better-informed purchase decisions.
2) Sustainability
   By reducing the need for excessive inventory and cutting down on returns, virtual try-on technology promotes a more sustainable retail model.
3) Customer Engagement
   The ability to virtually try on clothing enhances customer interaction and delivers a more engaging shopping experience.

Despite their transformative potential, building effective virtual try-on systems remains challenging. Accurate garment alignment, preservation of intricate details, and maintaining computational efficiency are all critical hurdles. These difficulties, combined with the rising demand for real-time, high-quality applications, drive the development of innovative solutions such as EfficientVITON.

1.2 Problem Statement

The objective of a virtual try-on system is to generate realistic images of a person wearing a target garment. Traditional systems rely heavily on paired datasets and external warping

networks to align garments with the human body. However, these approaches face significant limitations:

1) Generalizability [4]
   Conventional methods struggle to handle diverse body poses, complex backgrounds, and various clothing styles.
2) Loss of Detail
   Maintaining garment details such as textures, patterns, and fabric folds remains a challenge, especially in high-resolution applications.
3) Computational Overhead
   The iterative nature of many virtual try-on methods results in high computational demands, making real-time performance impractical.

Recent advancements in pretrained diffusion models, such as Stable Diffusion, have introduced robust generative frameworks capable of high-quality image synthesis. However, their application to virtual try-on tasks poses unique challenges [1][5][6]:

- Semantic Alignment
  Establishing precise correspondence between garments and body features.
- Processing Speed
  Addressing the inefficiency of standard diffusion processes without sacrificing quality.

EfficientVITON addresses these challenges by combining the strengths of diffusion models with innovative architectural enhancements, making it a highly efficient and scalable solution for virtual try-on.

1.3 Technical Foundations and Advancements

Virtual try-on systems rely on cutting-edge AI techniques to address the technical challenges of aligning garments, preserving details, and generating realistic outputs. Key technical aspects include [7] [8]:

1) Accurate Alignment with Human Bodies
   Virtual try-on requires precise spatial correspondence between garments and body features. This is achieved through:
   - Pose Estimation Models: Tools like OpenPose or DensePose estimate body keypoints and create detailed body maps, enabling structural alignment.
   - Warping Networks [9]: Learning spatial relationships between garments and body regions ensures accurate placement on diverse body types.
2) Preservation of Garment Details
   Capturing intricate details, such as patterns, textures, and folds, is critical. The solutions include:
   - High-Resolution Encoders: Spatial encoders designed to retain fine-grained features during image synthesis.
   - Attention Mechanisms: Advanced cross-attention blocks focus on integrating important garment features into the final output [7][1].
3) Computational Efficiency
   Real-time performance demands optimization of the generation process. Recent advancements include:
   - Non-Uniform Diffusion Timesteps: These reduce the computational burden of iterative processes by focusing on key timesteps, maintaining quality while improving speed.
   - Latent Space Operations: Diffusion within compressed latent spaces reduces memory requirements and speeds up generation.
4) Handling Complex Backgrounds and Input Variability
   Virtual try-on systems must perform well in diverse real-world settings:
   - Human Parsing Models: Segmenting images into body parts and clothing areas ensures accurate garment integration.
   - Agnostic Representations: Preprocessing inputs with agnostic maps and masks removes distractions from worn garments, focusing the model's attention on the new clothing.

EfficientVITON incorporates these advancements, leveraging the strengths of pre-trained diffusion models and introducing innovations like zero cross-attention blocks and spatial encoding [10]. These ensure:

- Precise garment alignment.
- Efficient computation.
- Preservation of high-fidelity details.

The combination of these technical foundations positions EfficientVITON as a state-of-the-art solution for virtual try-on systems, meeting the demands of both accuracy and efficiency in practical applications.

1.4 Applications and Use Cases

The versatility and practicality of virtual try-on technology have led to its adoption across various domains:

1) E-Commerce Integration
   Virtual try-on systems empower online retailers to provide personalized recommendations, reduce return rates, and improve customer satisfaction. Users can try garments digitally before making a purchase, fostering confidence in their choices.
2) Physical Retail Stores
   By integrating virtual try-on platforms in physical stores, businesses can offer virtual fitting rooms, enhancing the shopping experience and reducing reliance on traditional trial areas.
3) Personal Styling and Fashion Exploration
   Individuals can use virtual try-on systems to experiment with styles, colors, and fits, enabling more informed fashion decisions.
4) Sustainability in Fashion
   By reducing the need for excessive inventory and minimizing waste from returns, virtual try-on technology contributes to the environmental sustainability of the fashion industry.
5) Entertainment and Media
   Virtual try-on technology has applications in digital avatars, fashion shows, and media productions, allowing creators to visualize costumes and designs efficiently.

EfficientVITON's focus on speed and accuracy makes it suitable for real-time applications, setting the stage for widespread adoption in these areas.

1.5 Contribution

EfficientVITON represents a significant leap forward in virtual try-on technology, addressing the limitations of existing methods and introducing several key innovations:

1) Non-Uniform Diffusion Timesteps: By reducing the number of diffusion steps and employing a non-uniform distribution, our approach significantly reduces computational overhead while maintaining high output quality. This work is among the first to address the efficiency of a virtual try-on process through optimization of diffusion steps.
2) Zero Cross-Attention Blocks: These blocks establish semantic correspondence between garments and human bodies, enabling precise alignment without the need for external warping networks.
3) Spatial Encoder for Clothing Details: Our model incorporates a spatial encoder to preserve intricate details of garments, such as patterns and textures, ensuring high fidelity in the output.
4) Attention Total Variation Loss: This novel loss function enhances the sharpness and accuracy of attention maps, improving the realism of garment placement.
5) Practical Application: We developed a web platform powered by EfficientVITON, enabling real-time virtual try-on for individuals, retailers, and businesses.

These contributions establish EfficientVITON as a state-of-the-art virtual try-on system, bridging the gap between research advancements and real-world usability.

1.6 Thesis Structure

This thesis is organized as follows:

1) Introduction: Presents the motivation, problem statement, applications, and contributions of the work.
2) Related Work: Reviews the state-of-the-art virtual try-on systems and highlights their limitations.
3) Methodology: Describes the architecture of EfficientVITON, including the use of Stable Diffusion, zero cross-attention blocks, and non-uniform diffusion steps.
4) Experimental Results: Presents quantitative and qualitative evaluations, comparing EfficientVITON with existing methods.
5) Discussion: Explores the implications of the proposed methodology, analyzing its strengths, limitations, and potential improvements. It includes a reflection on computational efficiency, generalizability, and the realism of the generated outputs compared to existing methods.
6) Conclusion and Future Work: Summarizes the contributions and outlines potential directions for future research.

2 Related Work

2.1 Introduction to Virtual Try-On Systems

Virtual try-on systems have become a cornerstone of contemporary e-commerce, allowing users to see how clothing items would look on them without the need for physical trials. Early approaches relied on 2D image warping and stitching

techniques, which often fell short when handling complex poses, textures, or lighting variations. The emergence of deep learning, particularly generative models, has transformed this field by enabling more realistic and scalable solutions. In recent years, advances in diffusion models have further pushed the limits of image synthesis, delivering exceptional quality and stability in virtual try-on applications [11][3].

2.2 *Evolution of Generative Models in Virtual Try-On Systems*

Particularly Generative Adversarial Networks (GANs), the evolution of generative models in Virtual Try-On systems has become a cornerstone of virtual try-on systems, due to their ability to generate highly realistic images. Despite their popularity, GANs face several challenges, including mode collapse, training instability, and difficulties in preserving fine details in high-resolution outputs. As an alternative, diffusion models have gained traction by employing iterative denoising processes to produce high-quality images with greater stability. Recent advancements have showcased the potential of diffusion models in addressing fashion-specific challenges, including their application in virtual try-on tasks [4][9].

2.3 *Recent Developments in Virtual Try-On Systems*

Over the past five years, virtual try-on systems have advanced significantly as researchers explore new models and techniques to overcome critical challenges, such as pose alignment, texture preservation, and image realism. Below is an overview of the key approaches, their contributions, and their limitations:

1) **Geometric Warping-Based Models**
   Early virtual try-on systems heavily relied on geometric warping techniques to align clothing items with the target body pose as shown in Fig. 1. These methods typically employed 2D image transformations, such as Thin Plate Spline (TPS), to warp garments onto the user's image. For instance, VITON [12] introduced a TPS-based warping approach that enhanced alignment but struggled to preserve fine details and textures. While computationally efficient, these methods often produced unrealistic results when dealing with complex poses, occlusions, or fabric deformations.
   Strengths: Low computational cost and straightforward implementation.
   Weaknesses: Poor handling of complex poses, textures, and overall realism.

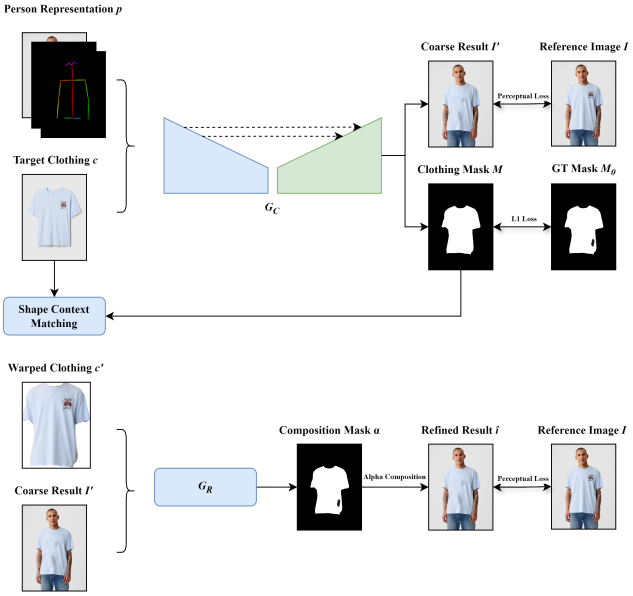

Fig. 1: Multi-Task Encoder Decoder Generator.

2) **GAN-Based Models**
   The advent of GANs revolutionized virtual try-on systems by enabling the realistic synthesis of garments on target poses. These methods generally operate in two stages: a warping module aligns the clothing with the body, followed by a refinement module that generates the final output. For example, VITON [12] and CP-VTON [13] are GAN-based frameworks that combined pose estimation with texture preservation to achieve state-of-the-art results as shown in Fig 2. However, GANs often face issues such as training instability, visible artifacts, and challenges in preserving fine details.
   Strengths: High-quality realism and the ability to handle intricate textures.
   Weaknesses: Unstable training, artifact generation, and difficulty in preserving details.
3) **Attention-Based Models**
   To address GANs' limitations, attention mechanisms were introduced to enhance alignment and texture preservation. These models use attention maps to focus on relevant regions of the clothing and the body, enabling more precise alignment and detailed output. For example, CP-VTON+ [14] developed an attention-based try-on network that excelled at managing complex patterns and textures. However, attention-based models tend to be computationally intensive and require large datasets for

training.

Strengths: Better alignment and improved texture detail.

Weaknesses: High computational cost and heavy data requirements.

4) **Diffusion Models**

Recently, diffusion models have emerged as a powerful alternative to GANs in virtual try-on applications. Unlike GANs, these models iteratively denoise images, which enhances stability and output quality. They have proven highly effective at managing complex textures [11] and Stable-VITON [3] are diffusion-based models that demonstrate exceptional image quality and stability as shown in Fig. 3. However, their iterative nature makes them computationally expensive, which poses challenges for real-time applications.

Strengths: Exceptional image quality, stability, and detail preservation.

Weaknesses: High computational demands and slow inference speeds.

5) **Hybrid Models**

Hybrid models combine the strengths of multiple approaches, such as geometric warping, GANs, and attention mechanisms, to achieve better alignment, realism, and efficiency. For instance, HR-VITON [15] proposed a hybrid framework that integrates geometric warping for initial alignment with GAN-based refinement for final synthesis as shown in Fig. 4. While effective, hybrid models often require complex architectures and significant tuning.

Strengths: Balanced performance by leveraging multiple techniques.

Weaknesses: Complex architecture and intensive tuning requirements.

6) **Multi-Modal Models**

Advancements in multi-modal inputs, such as text descriptions and sketches, have enhanced user interaction and personalization in virtual try-on systems. For example, Text2Cloth [16] developed a system allowing users to describe clothing in natural language, which the model then synthesizes onto their image. Similarly, Sketch2TryOn [17] introduced a sketch-based interface for designing and visualizing custom clothing in real-time. While engaging, these methods require additional preprocessing and are computationally expensive.

Strengths: Greater personalization and user engagement.

Weaknesses: High computational cost and preprocessing demands.

2.4 Current State-of-the-Art in Virtual Try-On

Diffusion-based methods currently lead the field, generating high-resolution, photorealistic images that preserve intricate clothing details. Recent advancements include integrating pose estimation, attention mechanisms for alignment, and multi-modal inputs for personalization. However, challenges such as adapting to diverse body shapes, managing occlusions, and improving real-time performance remain unresolved.

2.5 Research Gaps and Limitations

Despite significant progress, several gaps persist in virtual try-on research. Current methods often focus solely on static images, underutilizing temporal data for video-based try-on systems [18]. Personalization, such as accommodating user preferences or diverse body types, remains limited. The

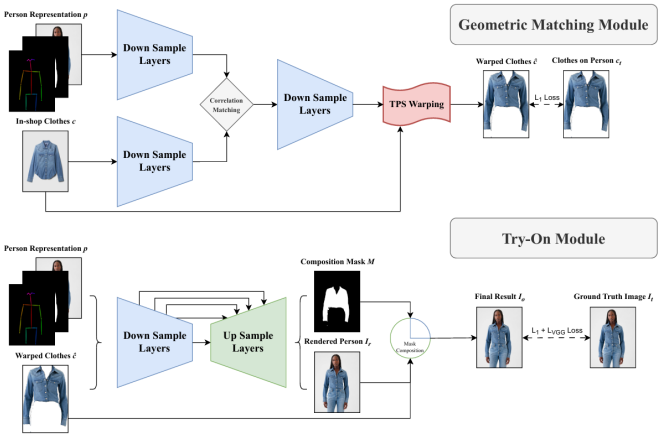

Fig. 2: VITON Model Architecture.

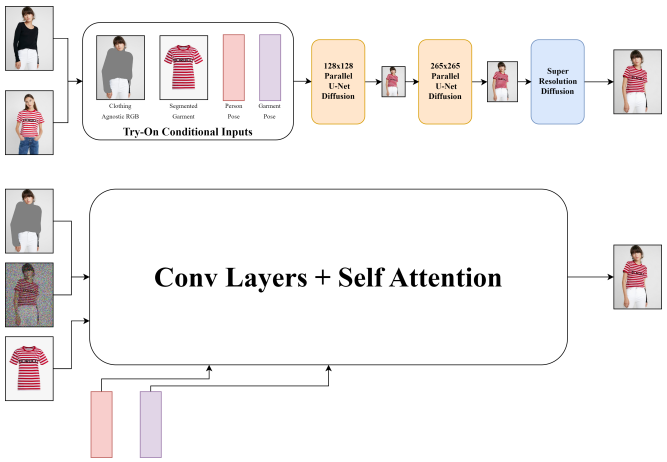

Fig. 3: TryOnDiffusion Model Architecture.

computational cost of diffusion models also restricts their use in real-time applications. Furthermore, there is a lack of standardized evaluation metrics tailored to virtual try-on systems, complicating objective comparisons between methods [14].

2.6 *Critical Analysis of Existing Work*

While existing systems have improved realism and usability, many struggle to generalize across diverse datasets and real-world scenarios. For instance, GAN-based methods often generate artifacts in complex scenarios, while diffusion models, despite their stability, remain computationally expensive [11][3]. Additionally, most systems rely on paired datasets, limiting scalability.

This research builds on recent advancements in diffusion models to address existing virtual try-on limitations. The proposed framework optimizes the diffusion process using non-uniform timesteps, reducing computational demands without compromising image quality. Incorporating temporal coherence for video-based try-on and a personalization module to adapt to individual user preferences and body shapes, this approach aims to bridge the gap between state-of-the-art methods and practical applications.

3 Methodology

3.1 *Data & Preprocessing*

For EfficientVITON to work as intended, the source image undergoes several preprocessing steps in order to help the model achieve the projected quality. These steps involve telling the model many useful information, like details about the posture of the person wearing the garment, abstracted image of the person without the initial cloth of interest, and identification of different human parts of the image.

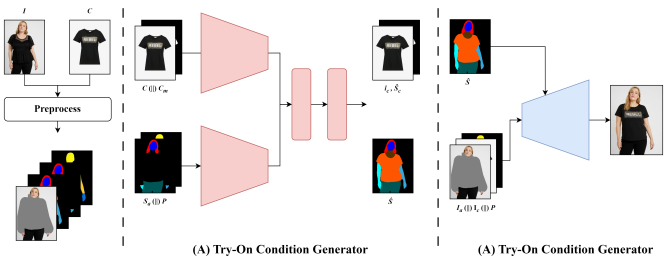

Fig. 4: HR-Viton Model Architecture.

Some of these steps are fed directly to the main model, and others are incorporated into other steps to produce an input for the model. The model architecture is assembled to make use of these inputs instead of the original unprocessed image to achieve higher accuracies and to discard any unnecessary details that will make the virtual try-on process harder and less efficient. For a virtual try-on model to work, it needs two main inputs, the image of the garment that the person wants to wear, and the image of this person wearing another garment of the same type. So, if they need to virtually try-on a top, their image should contain them wearing a top. These two main images undergo 8 different preprocessing steps:

1) Pose Estimation

The pose of the person is extracted into 25 keypoints and used to determine specific parts indicating the place of the worn garment. The 25 keypoints are distributed as follows:

- Head: Nose, neck, eyes, and ears.
- Upper body: Shoulders, elbows, and wrists.
- Lower body: Hips, knees, and ankles.

For estimating the human pose, we are using OpenPose for this task. OpenPose [19] [20] [21] [22] is a real-time multi-person state-of-the-art model for human pose estimation. It is the most efficient and precise model for this task. The model architecture shown in Fig. 6 consists of two main parts:

(A) Convolution Neural Network (CNN) Model: The input image is passed through multiple convolution layers to extract feature maps from the image. The CNN is based on a pre-trained network, like VGG-19.

(B) Heatmaps: The feature maps are then used to generate

Fig. 5: OpenPose Output.

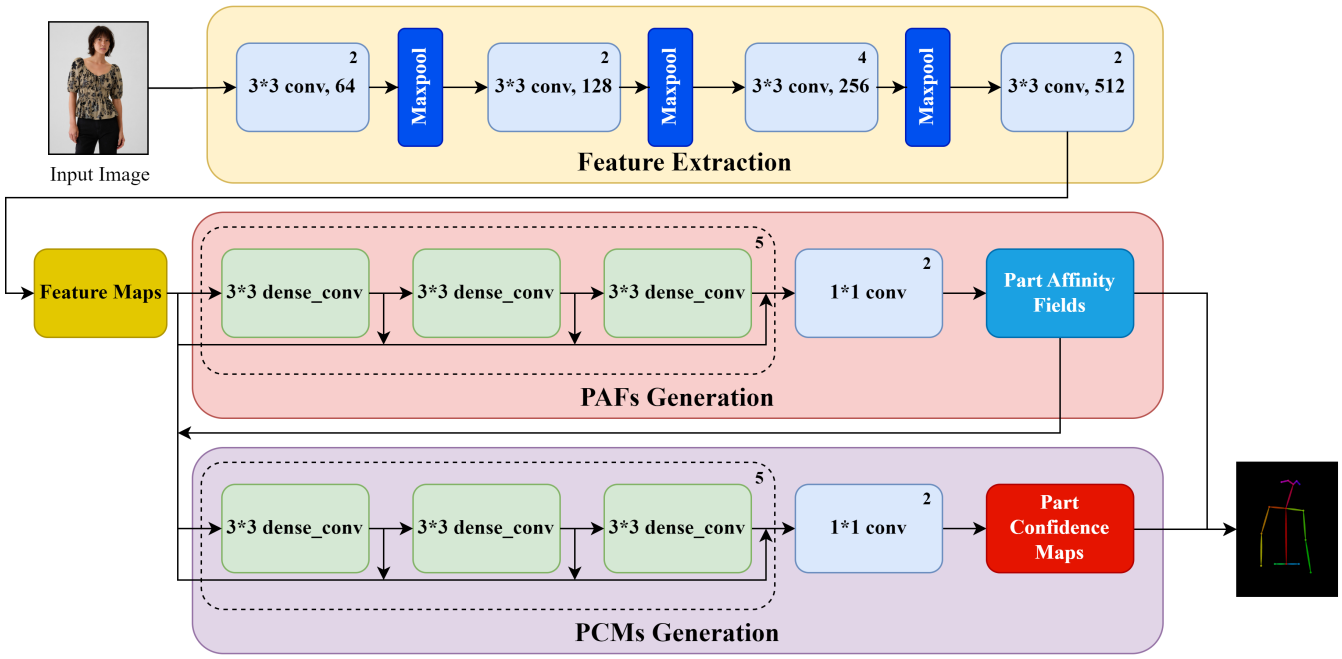

Fig. 6: OpenPose Model Architecture.

two sets of heatmaps, one for Part Confidence Maps (PCMs), and the other for Part Affinity Fields (PAFs). The PCMs are used to indicate the probability of a body part existing in a location in the image. The PAFs are used to indicate the directional connections between these body parts. Utilizing both PAFs and PCMs, OpenPose uses a greedy bipartite matching algorithm to generate the pose keypoints and the connections between them.

The OpenPose output shown in Fig. 5 is used to determine the places of the head and upper body areas. These will be used in extracting the Agnostic Image and the Agnostic Mask later in steps 3 and 4.

2) Human Parsing

The human image is parsed into 20 different classes according to the body parts and worn clothes and used alongside the output of OpenPose to determine the place of the worn garment. The 20 classes are: Background, Hat, Hair, Glove, Sunglasses, Upper-clothes, Dress, Coat, Socks, Pants, Jumpsuits, Scarf, Skirt, Face, Left-arm, Right-arm, Left-leg, Right-leg, Left-shoe, and Right-shoe. For the human parsing task, we are using the Look Into Person (LIP) Parsing Model. LIP Parsing [23] is based on semantic segmentation to parse the human image into fine-grained body regions. Like any other semantic segmentation model, it assigns a class for each pixel in the image from the 20 mentioned classes, segmenting the human image into different semantic regions of clothing and body parts.

The backbone of the LIP Parsing model architecture, as shown in Fig. 7, is a pretrained CNN based on ResNet to extract features from the image. These features are passed to an Atrous Spatial Pyramid Pooling (ASPP) block to extract multiscale contextual information by using multiple parallel CNNs with different dilation rates to capture features at various scales. This is the main segmentation part in the architecture. The output of the ASPP block is fed into a decoder network which generates high-resolution segmentation maps. As shown in Fig. 8, the output of the LIP Parsing model is used alongside the OpenPose output to generate the Agnostic Image and the Agnostic Mask in steps 3 and 4. It is also used to generate

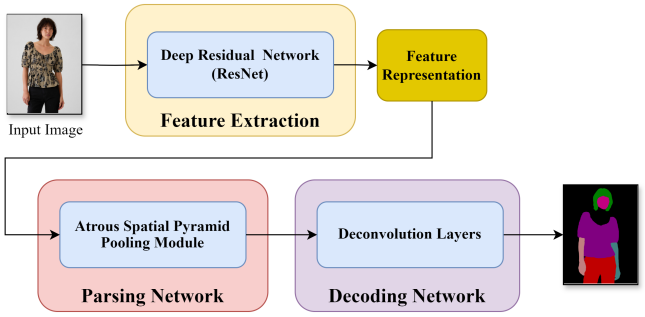

Fig. 7: LIP Parsing Model Architecture.

Fig. 8: LIP Parsing Output.

the Ground Truth Warp Mask in step 6.

3) **Agnostic Image**

The agnostic image shown in Fig. 9 is the human image but abstracted from the target clothing area. For example, if the person is virtually trying-on a top, the agnostic image will be abstracted from the whole top area, neck to hands to hips. This is to help the model apply the new garment in this area without the noise of the already worn garments.

The algorithm for generating the agnostic image takes in the pose keypoints from the OpenPose Model and the human parsing map from the LIP Parsing Model to generate the agnostic image. It uses these inputs to draw a grey mask covering the whole area of the target clothing.

4) **Agnostic Mask**

The agnostic mask shown in Fig. 10 is the mask of the target clothing area. It undergoes the same steps of

Fig. 10: Agnostic Mask Output.

generation of the agnostic image, but the grey mask is extracted into another image instead of applying it on the same image.

While the agnostic image guides the model for the input image without any noise from the already worn garments, the agnostic mask guides the model to the area that it should fill with the new garment.

5) **Parse Agnostic**

The parse agnostic image shown in Fig. 11 uses the human parsing semantic map and the extracted agnostic mask to remove the agnostic mask area from the semantic map. The output image will be the agnostic mask subtracted from the human parse semantic map. This, alongside the agnostic mask, helps the model to determine exactly the parts it needs to fill during the image generation process, and the parts the it needs to keep.

6) **Ground Truth Warp Mask**

The ground truth warping mask, shown in Fig. 12 is the mask of the original garment worn by the person. This

Fig. 9: Agnostic Image Output.

Fig. 11: Parse Agnostic Image Output.

Fig. 12: Ground Truth Warp Mask Output.

Fig. 14: DensePose Output.

is used during the training process to adjust the loss of the model. The loss measures the difference between the generated garment area and the already worn area, so the model tries to minimize this difference. The algorithm for extracting the ground truth warping mask relies on the output of the LIP Parsing model by selecting multiple classes and extracting them from the image to indicate the mask of the already worn garments.

7) **Dense Human Pose Estimation**

A dense human pose is a 3D surface model of the body. The 3D surface model will preserve the details of the textures and the pose of the person during the virtual try-on process. For this task, the DensePose model was used.

DensePose [18] is a state-of-the-art model for the 3D surface model generation from the 2D image. Built on Mask R-CNN framework, DensePose's architecture, shown in Fig. 13, extends it by adding a dense regression branch that predicts UV coordinates for each pixel. The UV coordinates are used to establish a relation between each pixel in the original 2D image and a specific point on the 3D model of the human body.

As shown in Fig. 14, the output of the DensePose Model

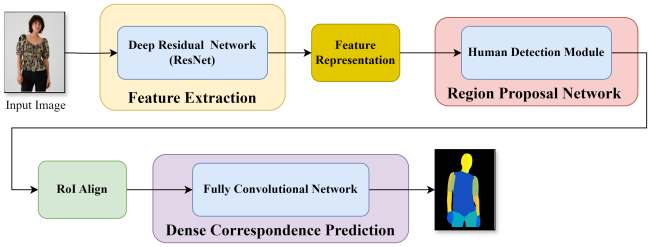

Fig. 13: DensePose Model Architecture.

helps the model preserve the texture and the pose in the generated part which is covered by the agnostic mask. For example, if the person is virtually trying-on a summer top, parts of the arms will be covered by the agnostic mask, but will need to be generated again during the image generation process. To preserve the body pose and shape, and enhance realism, EfficientVITON uses the DensePose map to accurately re-generate the masked parts.

8) **Cloth Mask**

This is the only preprocessing part that is done on the garment image solely, not on the person image. The garment image undergoes a semantic segmentation process to divide the image into two classes: background and garment. The garment is the class of interest, so its mask is extracted. The cloth mask in Fig. 15 helps the model indicate exactly where is the garment that is supposed to use in the virtual try-on process. This improves the generalizability of cloth inputs to the model, allowing different cloth from different environments to be used as an input to the model without the problem of keeping a standard.

Fig. 15: Cloth Mask Output.

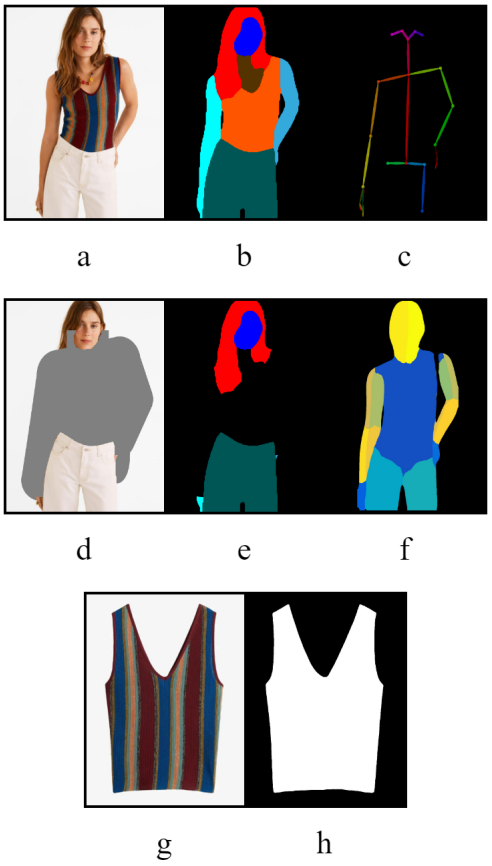

Fig. 16: A sample from VITON-HD dataset. a) Normal Person Image. b) Human Parsing. c) OpenPose Pose Estimation. d) Agnostic Image. e) Parse Agnostic Image. f) DensePose Image. g) Unworn Cloth Image. h) Cloth Mask.

Training Dataset

The main training dataset used was VITON-HD [1]. This dataset was collected by Zalando, a German online retailer, in contribution for the research in one of the first virtual try-on researches by Choi et al. [1]. The dataset contains over 11600 training images and over 2000 testing images. All images are 1024x768 in terms of resolution. Each image in the dataset has its equivalent image of the unworn clothes. In addition, the dataset handles most of the preprocessing steps, like the densepose, the human parsing, the pose estimation, etc. Fig. 16 shows a sample from the training dataset with the available types of preprocessing.

3.2 Stable Diffusion Architecture

Stable Diffusion, the core generative model architecture of EfficientVITON, is renowned for its ability to synthesize high-quality images efficiently within a compressed latent space [24]. This section elaborates more on Stable Diffusion's workings and justifies its selection as the foundation for the virtual try-on task. Fig. 17 visually represents the Stable Diffusion architecture.

Stable Diffusion is built upon three key components:

1) Variational Autoencoder (VAE)

The VAE acts as a bridge between the high-dimensional pixel space and a lower-dimensional latent space, crucial for computational efficiency. It comprises two parts:

- **Encoder (E):**
  This component maps an input image $x \in \mathbb{R}^{H\times W\times 3}$ to a compressed latent representation $z = E(x) \in \mathbb{R}^{h\times w\times c}$, where $h = H/f$, $w = W/f$, and f is the downsampling factor. This compression reduces the computational complexity of the subsequent diffusion process. The encoder strives to preserve essential image information within this compressed representation.
- **Decoder (D):**
  The decoder reconstructs the image $\hat{x} = D(z)$ from the latent representation z. The VAE is trained to minimize the difference between the original image x and the reconstructed image $\hat{x}$. This difference is typically measured using a reconstruction loss, such as the mean squared error (MSE):

$$\mathcal{L}_{rec} = ||x - D(E(x))||^2 \tag{1}$$

In addition to the reconstruction loss, a regularization term, often the Kullback-Leibler (KL) divergence, is used to ensure that the learned latent space adheres to a prior distribution, typically a standard Gaussian $\mathcal{N}(0, I)$. The overall VAE loss is a weighted combination of these two terms:

$$\mathcal{L}_{VAE} = \mathcal{L}_{rec} + \beta\mathcal{L}_{KL} \tag{2}$$

where β is a hyperparameter controlling the strength of the KL regularization.

2) U-Net Denoising Network

The U-Net is a convolutional neural network specialized for image-to-image translation tasks, and it's the heart of the denoising diffusion process. The network takes a noisy image in the latent space and attempts to predict the noise added to the original latent representation. It operates iteratively, progressively removing noise to refine the generated image.

The Key U-Net features are:

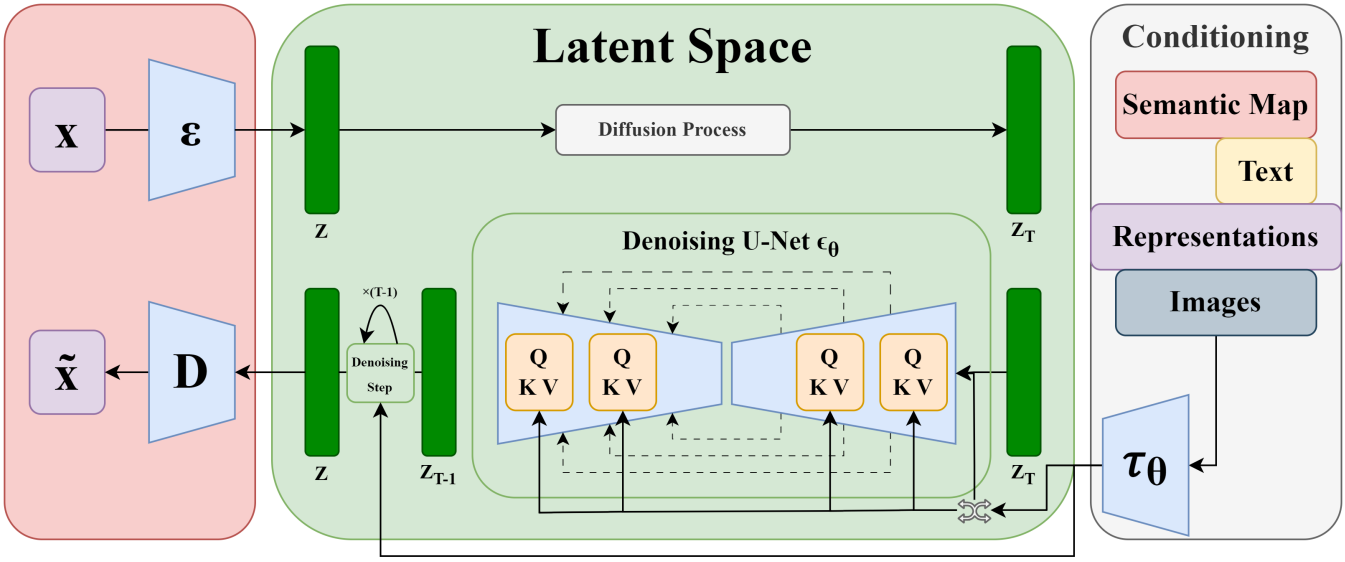

Fig. 17: Stable Diffusion Architecture

(A) **Symmetric Encoder-Decoder Architecture with Skip Connections:**
Skip connections directly link corresponding layers in the encoder and decoder pathways. This architectural design allows the network to bypass the bottleneck layers, facilitating the preservation of fine details and structural information from the input during the denoising process.

(B) **Residual Blocks:**
The incorporation of residual blocks aids in stabilizing the training process, especially in deep networks. These blocks allow gradients to flow more easily during backpropagation, mitigating issues like vanishing gradients.

(C) **Multi-Head Attention Layers:**
Attention mechanisms, and particularly multi-head attention, enable the network to weigh different parts of the input image differently. This selective attention allows the network to focus on crucial regions during each step of the denoising process, adapting its behavior based on the context.

3) Diffusion Process (Denoising)

Diffusion models operate by progressively adding and removing noise from an image. The process can be divided into two phases:

(A) **Forward Diffusion:**
This process gradually adds Gaussian noise to a latent representation z_0 over a series of timesteps $t = 1, ..., T$, following a predefined variance schedule β_t. This process continues until the latent representation is essentially pure noise.

(B) **Reverse Diffusion (Denoising):**
This process, learned by the U-Net, attempts to reverse the forward diffusion process. At each timestep, the U-Net, parameterized by θ, predicts the distribution of the previous latent state z_{t-1} given the current noisy latent z_t.

Stable Diffusion performs this iterative denoising process within the VAE's latent space, offering significant computational and memory advantages over pixel-based diffusion.

Stable Diffusion's benefits for EfficientVITON virtual try-on task are threefold:

- **High-Fidelity Generation**
  Stable Diffusion is capable of generating high-quality, realistic images with fine-grained details—essential for convincingly visualizing garments on individuals in a virtual try-on scenario.
- **Efficiency**
  Operating in the compressed latent space drastically reduces memory and computational demands, making it feasible to generate high-resolution images essential for capturing intricate clothing details.
- **Pretraining**
  Stable Diffusion models are typically pretrained on vast image datasets, such as LAION [2]. This provides the model with a strong prior understanding of diverse image

content, including human poses, textures, and clothing styles, enabling effective generalization and faster fine-tuning for the specific task of virtual try-on.

3.3 EfficientVITON Architecture

The model's goal is to perform image-based virtual try-ons, creating a realistic image of a person wearing a specific clothing item. Unlike previous methods, which frequently struggle with generalizability and preserving fine details, particularly in complex backgrounds, EfficientVITON leverages the power of pretrained diffusion models, specifically the Stable Diffusion Model, as previously discussed. This enables it to inherit strong generative capabilities, resulting in more natural and visually appealing try-on outcomes. The key innovation is the model's ability to learn the semantic correspondence between the clothing item and the human body within the latent space of the pretrained diffusion model. This direct learning of relationships within the latent space, rather than relying on external warping networks, is a critical differentiator that allows the model to maintain clothing details while benefiting from the pretrained model's generative power.

3.3.1 Inputs

The model requires several input conditions to perform the virtual try-on task:

- Clothing Image (x_c): An RGB image (3 channels) of the clothing item to be virtually tried on. Similar to the person image, higher resolution inputs allow for finer details in the generated output.
- Agnostic Map (x_a): A modified version of the person image where the original clothing has been removed or masked. This is created during preprocessing (step 3) by inpainting the region defined by the pose estimation and human parsing results (steps 1 and 2), typically replacing the clothing area with neutral grey tones. The agnostic map maintains the same resolution and three color channels as the original person image. This input helps the model focus on integrating the new clothing item seamlessly onto the person.
- Agnostic Mask (x_{ma}): A single-channel binary mask that corresponds to the regions of the original clothing in the person image. Generated in preprocessing (step 4), this mask, which is having the same resolution as the person image, guides the model to focus its generation efforts on the areas where the new clothing should be placed, helping to avoid artifacts or inconsistencies.
- Dense Pose (x_p): This input provides detailed 3D surface information about the person's pose. Generated using DensePose in the preprocessing (step 7), based on the Mask R-CNN framework, it maps each pixel of the person image to UV coordinates on a canonical 3D human model. The output of DensePose is a UV map, which is then encoded into a latent representation before being input to the model. This latent representation of the dense pose helps the model preserve the person's body shape and pose during the virtual try-on, resulting in more realistic and natural-looking clothing drape.

3.3.2 Architecture

In this part, different parts of the model's architecture [3] (Fig. 18) are discussed. Starting from our model's base, the pretrained Stable Diffusion Model, handling the main diffusion process. It is paired with a spatial encoder to preserve the clothing details. The decoder part is injected with nine Zero Cross-Attention Blocks to combine the clothing item with the input images. All of these parts are combined together to deliver a high-fidelity virtual try-on image.

Usage of Stable Diffusion Model

The core of EfficientVITON architecture is the pretrained Stable Diffusion Model. Our model leverages the superiority of the Stable Diffusion Model in the diffusion process in order to generate high-quality images. The Stable Diffusion Model is pretrained on LAION (Large-scale Artificial Intelligence Open Network) dataset [2]. The model's state-of-the-art performance and the quality of the output images are the main reason we chose it as our main image-to-image translation medium.

The utilization of the pretrained Stable Diffusion Model allows a lot of advantages and benefits for the architecture of EfficientVITON model. First, the model's ability to generate high-quality images provides a steady base and a robust foundation for realistic virtual try-on results. Second, using a pretrained diffusion model allows the resulted virtual try-on model to make use of a wide range of realistic data including complex features of humans and clothes. Thus, making the development of the virtual try-on model faster without the need to relearn basic features and only focusing on the main process, primarily learning the semantic correspondence between clothing and the human body. In addition, the high availability and the flexibility of Stable Diffusion Model allows various integrations in different parts of the main architecture, which gives a wide range of capabilities for our research.

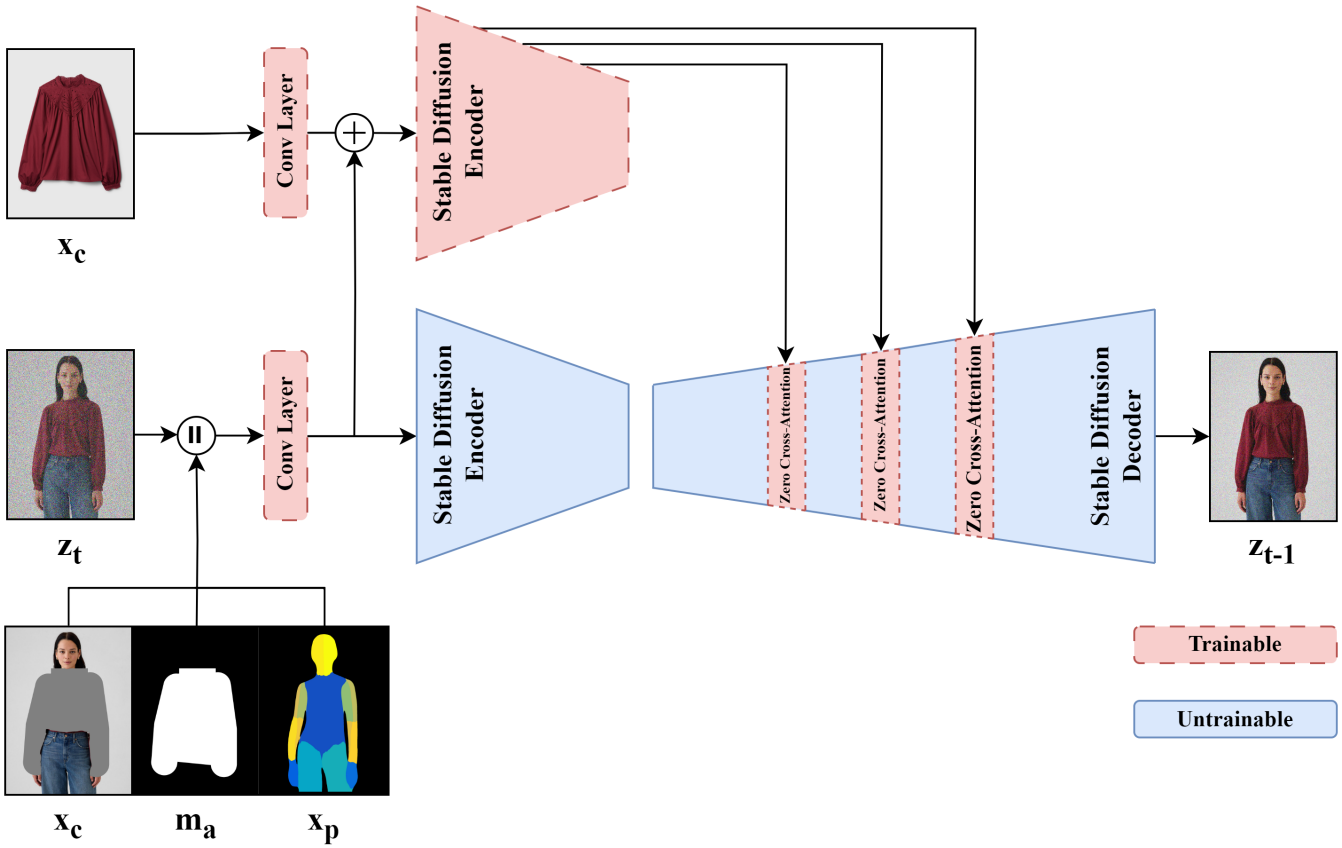

Fig. 18: EfficientVITON Model Architecture.

While the pretrained Stable Diffusion Model is our basis, it will interact with other components of the architecture, such as the spatial encoder and zero cross-attention block, in order to perform the virtual try-on process. The latent maps are manipulated by these components which leads to a transformation of the clothing with respect to the target person. Then, the U-Net which is responsible for denoising the latent map receives all the inputs (latent map, clothing information, dense pose etc.) to perform the required operation. The decoder of the Stable Diffusion Model is supposed to translate the latent maps into the final output of an image with the new cloth worn on the person. In conclusion, the Stable Diffusion pretrained model provides the essential architecture upon which all other processing and generation is based.

Spatial Encoder

For a virtual try-on model to produce high quality images, it should preserve the fine details of the clothing, like the colors and textures on the clothing. That's where the role of the spatial encoder comes in. It is designed to thoroughly preserve the fine-grained details of the input clothing item. Previous virtual try-on methods [15][11][1] struggle with clothing with complex textures and high frequency features, resulting in a loss of realism and producing low quality virtual try-on results. To address this limitation, the spatial encoder explicitly captures the spatial information of the clothing image, deliberately passing this information into the virtual try-on process, and ensuring that the clothing features are accurately transferred to the output virtual try-on image. The spatial encoder's primary aim is to inject intermediate features of the clothing to the U-Net via the zero cross-attention block and ensure that no information from the clothing is lost during this process.

The spatial encoder is designed to work in parallel with the encoder part of the Stable Diffusion U-Net. It performs the same process but on the clothing item, employing the same sequence of downsampling and convolutional layers. This design choice is strategically intentional, as both encoders share the same weights at the start of the training. This allows the spatial encoder to benefit from the feature extraction layers of the pretrained U-Net of the Stable Diffusion's encoder. This allows the virtual try-on model to benefit from the underlying pretrained layers while focusing only on the main virtual try-on task and avoiding the need to train a new encoder from

scratch.

The clothing item is fed to this encoder input which translates it into a set of intermediate feature maps at multiple resolutions. The output of the spatial encoder is fed to the key (K) and value (V) inputs of the zero-attention blocks which learns the semantic correspondence with respect to the spatial features of the clothing and the query input. By integrating the clothing item feature maps into these blocks and into the decoding parts, the model is learning more accurately about the semantic correspondence between the clothing and the person. Additionally, it allows the model to preserve the clothing details during the denoising process in the decoder, which produces more realistic images in terms of clothing details.

In summary, the spatial encoder plays a crucial role in ensuring detailed clothing preservation within the virtual try-on framework. It complements the other modules of the architecture, providing them with spatially refined information that is vital for accurate clothing alignment and high-quality image generation. This ensures that the textures, patterns, and overall appearance of the clothing are accurately represented, contributing to the overall realism of the generated virtual try-on image. Therefore, the combination of spatial encoder and cross-attention blocks results in state-of-the-art results for virtual try-on.

Zero Cross-Attention Blocks

Previous versions of the virtual try-on models often struggle with deforming and aligning the clothing item to fit the person body and pose accurately, which results in unnatural or misaligned results. Our architecture addresses this problem by introducing the zero cross-attention blocks which are designed specifically to learn the semantic correspondence between the clothing item and the human body and pose. These blocks are injected into the pretrained diffusion model and act as the channel between the clothing and the human body.

The core of the zero cross-attention block (Fig. 19) lies in its usage of the attention mechanism to learn this interaction between the human body regions and the clothing item spatial features. Specifically, the query (Q) part of the cross-attention mechanism takes input from the feature maps generated by the U-Net's decoder from the pretrained Stable Diffusion Model. These feature maps represent the human body areas where the clothing should be applied. On the other side, the key (K) and the value (V) parts take input from the spatial encoder, which represents the features of the clothing item. By computing attention weights through the dot product between key and query, the model establishes relationships between the relevant regions of the clothing and the corresponding regions of the human body, enabling precise alignment of clothing details with the target person.

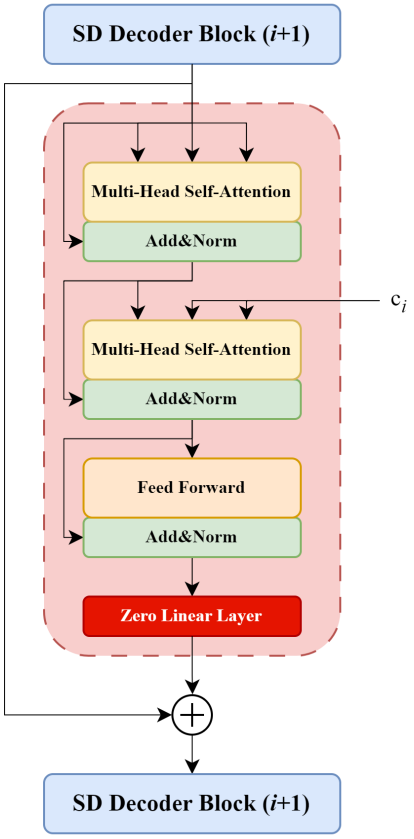

Fig. 19: Zero Cross-Attention Block.

This process allows for a patch-wise warping of the clothing item on the human body within the latent space of the model. To eliminate any kind of noise artifacts, the result of the cross-attention part is passed through a linear layer, initialized with zero weights.

Architecturally, the position of the zero cross-attention blocks between the decoder layers of the pretrained Stable Diffusion Model allows a seamless integration of clothing item spatial information throughout the image generation process. As they take input from the spatial encoder which provides feature maps of the clothing item, they provide the decoder with the clothing information, while the person information is provided from the pretrained encoder. Injecting these blocks between the decoder layers makes sure that each layer takes into account the provided clothing item information.

3.3.3 Loss Functions

The training process of the model makes use of a combination of loss functions to achieve optimal performance. The primary loss function is inherited from the pretrained Stable Diffusion model [24], while a novel loss is introduced to refine the attention mechanism for accurate clothing placement.

Latent Diffusion Model Loss (L_{LDM})

The primary loss function, L_{LDM}, is derived from the pre-trained Stable Diffusion model. This loss guides the U-Net's learning of the denoising process within the latent space, aiming to minimize the difference between added and predicted noise:

$$L_{LDM} = \mathbb{E}_{\zeta, x_c, \epsilon, t} \|\epsilon - \epsilon_\theta(\zeta, t, T_\phi(x_c), E(x_c))\|^2 \quad (3)$$

Where:

- ζ: Concatenated latent inputs (noisy latent map z_t, latent agnostic map $E(x_a)$, agnostic mask x_{ma}, latent dense pose $E(x_p)$).
- x_c: Clothing image.
- ϵ: Random noise.
- ϵ_θ: U-Net parameterized by θ.
- t: Timestep.
- T_ϕ: CLIP (Contrastive Language-Image Pretraining) image encoder parameterized by ϕ.
- E: VAE encoder.

Attention Total Variation Loss (L_{ATV})

To enhance attention map quality, EfficientVITON utilizes the Attention Total Variation Loss (L_{ATV}):

$$L_{ATV} = \|\nabla(FM)\|_1 \quad (4)$$

Where:

- F: Center coordinate map (derived from attention maps).
- M: Ground truth clothing mask.
- ∇: Gradient operator.

Overall Loss

The overall loss function balances these components:

$$L_{total} = L_{LDM} + \lambda_{ATV} L_{ATV} \quad (5)$$

Where λ_{ATV} is a hyperparameter weighting the L_{ATV} contribution.

3.3.4 Training Process

The training of the model is a two-stage fine-tuning process designed to leverage the pretrained Stable Diffusion model effectively while adapting it for the virtual try-on task.

Stage 1: Semantic Correspondence Learning

In the first stage, the primary focus is on establishing robust semantic correspondences between the clothing item and the person's body. This stage fine-tunes the U-Net, specifically the zero cross-attention blocks, to learn these relationships within the latent space. Crucially, data augmentation plays a vital role in this stage. The input clothing and person images are augmented using techniques like random shifts, random scaling, horizontal flips, and color jittering. These augmentations force the model to learn correspondences that are invariant to these transformations, resulting in more robust and generalized mappings between clothing features and body regions. The model is trained using the standard latent diffusion loss (LLDM) during this stage. The AdamW optimizer is used with a learning rate of 1e-4 for 360,000 iterations with a batch size of 4.

Stage 2: Attention Map Refinement

The second stage builds upon the learned semantic correspondences by refining the attention maps generated by the zero cross-attention blocks. This stage introduces the attention total variation loss (LATV), which encourages sharper and more concentrated attention maps. This refinement leads to more accurate clothing placement and reduces artifacts or blurring around the clothing boundaries. The model is fine-tuned using both LLDM and LATV, with a weighting hyperparameter λ_{ATV} balancing their contributions. This stage uses the same optimizer and batch size as stage 1, but with a reduced number of iterations.

3.3.5 Inference Process

Once trained, the model generates virtual try-on images through the following inference process:

Latent Space Sampling:

The preprocessed input images (clothing image, agnostic map, agnostic mask, and dense pose) are encoded into the latent space using the Stable Diffusion VAE. The U-Net, conditioned on the encoded clothing and person information, then performs a reverse diffusion process, iteratively denoising a

random latent vector to generate the final latent representation of the try-on image. The model uses the Pseudo Linear Multi-step (PLMS) sampler for this denoising process.

RePaint for Paired Evaluation:

When evaluating the model in paired settings (i.e., where the ground truth try-on image is available), the model employs the RePaint approach. This technique iteratively replaces the regions of the generated image corresponding to the agnostic map with the actual agnostic map from the input during the sampling process. This helps preserve the details and context of the original person image in the non-clothing areas, leading to more accurate and visually appealing results for evaluation purposes.

Decoding to Image Space:

The final latent representation is decoded back into the pixel space using the Stable Diffusion VAE decoder, resulting in the final virtual try-on image.

3.3.6 Summary of Key Contributions

EfficientVITON's state-of-the-art performance stems from the following key innovations:

- **End-to-End Virtual Try-On with a Pretrained Diffusion Model:** Unlike previous methods that rely on separate warping modules, the model performs the entire virtual try-on process within the framework of a pre-trained diffusion model, offering a more integrated and efficient approach.
- **Latent Space Semantic Correspondence Learning via Zero Cross-Attention Blocks:** The novel zero cross-attention blocks enable the model to learn the complex relationships between clothing and the human body directly within the latent space, resulting in more accurate and natural-looking try-on results.
- **Attention Total Variation Loss and Data Augmentation:** The introduction of the LATV loss, along with strategic data augmentation techniques, significantly improves the sharpness and accuracy of the learned attention maps, further enhancing the quality and realism of the generated virtual try-on images.

These contributions combine to make this model a powerful and effective approach for virtual try-on, holding significant promise for practical applications in various domains.

3.4 Efficient Diffusion Process

The diffusion process in EfficientVITON architecture depends on Stable Diffusion Model as the main image-to-image translation model. While the quality of the output of Stable Diffusion is relatively high, it is not efficient enough. It uses a lot of memory resources and needs a lot of time to complete the diffusion process. Jiang et al. addressed the problem of an efficient diffusion process in their research [25], in which we adopted a similar approach to this problem. The main solution to this problem is to modify the timesteps required for the model to do the denoising process. Instead of a large number of uniformly distributed timesteps, we apply a non-uniform distribution of a small number of timesteps in the denoising process.

Instead of sampling $\boldsymbol{n}$ steps uniformly from all possible timesteps, we sample from a smaller set of strategically chosen timesteps. This significantly reduces the time needed for the diffusion process without sacrificing the quality of the output image, as the model will now learn the most significant timesteps in the denoising process and try to take more impulsive steps towards the final goal. The non-uniform distribution allows the model to take different amounts of denoising steps according to the position of the timestep. Fig. 20 shows a description of the non-uniform denoising steps which allows the model to be more efficient.

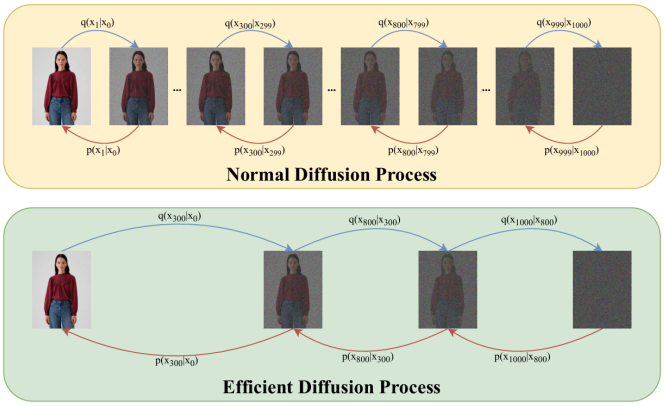

Fig. 20: Illustration of the normal diffusion process vs. EfficientVITON diffusion process.

4 Results

Our study yielded very encouraging results, demonstrating that our suggested model, EfficientVITON, can accurately align and smoothly adapt a person's appearance with the intended clothing. This development shows that virtual try-on

systems have advanced significantly, especially in producing realistic, high-quality results. Here, we outline the steps we took to guarantee that our model would successfully tackle the difficulties present in this field. In order to support our findings, we thoroughly compared EfficientVITON with a number of cutting-edge techniques in the virtual try-on space, utilizing both qualitative and quantitative analyses based on industry-accepted criteria.

Three virtual try-on models based on GANs (VITON-HD [1], HR-VITON [15], and GP-VTON [26]) and two diffusion-based models (LADI-VTON [9] and DC-I-VTON [4]) were included in our comparative analysis. Furthermore, as a baseline for evaluating inpainting quality in virtual try-on applications, we included Paint-by-Example [27], a diffusion-based inpainting technique. This thorough comparison shows how our model outperforms current methods in terms of both visual and quantitative performance.

We used the VITON-HD [1] dataset, which is publically accessible and commonly used as a standard in virtual try-on research, to assess EfficientVITON's performance. A realistic and demanding testbed for assessing virtual try-on systems is provided by the dataset's high-resolution photos. In particular, we employed this dataset's test set, making sure that our analyses followed established procedures to preserve the accuracy and validity of the findings.

The following three crucial dimensions were the focus of our evaluation:

1) Qualitative Results: Ensuring the visual realism and accuracy of the generated images was one of our main goals. The outputs' ability to smoothly incorporate the desired clothes with the subject's image was used to evaluate their visual appeal. In particular, we looked at elements like:
   a) Alignment: The precise placement of the garments on the intended body.
   b) Consistency: Maintaining the individual's natural characteristics, including posture and body composition, without adding any alterations.
   c) Detail Transfer: Preserving the clothing's fine features, such as its colors, textures, and patterns, without creating artifacts.

   The qualitative findings showed that EfficientVITON continuously produced more accurate and aesthetically pleasing outputs than other models.
2) Quantitative Results: We used two well-known metrics, Fréchet Inception Distance (FID) [28] and Learnt Perceptual Image Patch Similarity (LPIPS) [29], to objectively assess the performance of our model. Because of their capacity to capture many facets of the generated images:
   a) FID: Indicates how closely the distribution of generated and genuine images resembles one another; better realism is indicated by lower scores.
   b) LPIPS: Assesses perceptual similarity by contrasting the target and generated pictures' visual and structural coherence. Higher target image fidelity is indicated by lower LPIPS scores.
3) Efficiency Results: Making sure EfficientVITON could produce excellent results while preserving computing economy was a crucial component of our study. For real-world applications where resource limitations are a major factor, this is especially crucial.

Despite their transformative potential, building effective virtual try-on systems remains challenging. Accurate garment alignment, preservation of intricate details, and maintaining computational efficiency are critical barrier. These difficulties, combined with the growing demand for high-quality real-time applications, drive the development of innovative solutions such as EfficientVITON.

4.1 Qualitative Results

The visual quality of the outputs produced by our model (Fig. 21), EfficientVITON, shows how well it can handle the complex problems related to virtual try-on systems. As seen by the accompanying photographs, the model's ability to produce visually appealing outcomes is highlighted in this section. This corresponds with the intended apparel features and person-specific details.

Our model demonstrates an impressive ability to precisely represent the features of a wide range of apparel items, including items with intricate patterns, text, and images. For example, apparel with text or logos is rendered with remarkable clarity and alignment, guaranteeing that these tiny details are readable and aesthetically consistent in the finished product. This level of details demonstrates how effectively our method works to preserve the integrity of complex design features, which are frequently subjected to distortion in virtual try-on systems.

Additionally, EfficientVITON exhibits its versatility by effectively fitting clothing onto people with a wide range of skin

Fig. 21: Qualitative Results of EfficientVITON for the VITON-HD Test Dataset.

tones, facial features, and body shapes. This range of representation highlights how well the model generalizes, eliminating biases and guaranteeing accurate results for all demographic groups. The outputs highlight the model's realism and personalization potential by preserving the people' inherent features, including facial structure and body proportions, without the introduction of artifacts or distortions.

Adapting to clothing items with different levels of coverage is another strength of our model. The model maintains seamless transitions and guarantees correct alignment with the subject's body contours when wearing clothing that covers bigger areas of the body, like long-sleeved shirts. On the other hand, EfficientVITON precisely integrates the clothing with the exposed areas of clothing that exposes more skin, like tank tops or short-sleeved shirts, resulting in outputs that are both aesthetically beautiful and naturally occurring. This adaptability shows that the model can manage a variety of clothing designs without compromising the quality of the overall visual product.

The outcomes also demonstrate how well EfficientVITON handles alignment issues, in which the clothing needs to fit the subject's posture and body type. The produced images exhibit exact alignment, preventing problems like garment misplacement or abnormal warping. The outputs preserve the clothing's natural folds, textures, and shadows, adding to the impression of authenticity and guaranteeing that they cannot be visually distinguished from actual photos. More image results are shown in Fig. 22 in Appendix A.

EfficientVITON's visual appearance raises expectations for virtual try-on systems. The model's extensive skills and robustness are demonstrated by its ability to handle a variety of clothing designs, adjust to different subject features, and maintain realism across varying clothing coverage levels. These accomplishments establish EfficientVITON as an innovative approach for producing excellent, aesthetically pleasing results, opening the door for more advancements in this field.

4.2 Quantitative Results

A quantitative analysis was performed utilizing two commonly used measures, FID [28] and LPIPS [29], in order to

thoroughly assess the effectiveness of EfficientVITON. These measures were selected to evaluate the generated images' realism and perceptual quality, offering a strict standard by which to compare them to the most advanced virtual try-on models now in use. (Table I) provides a summary of the findings from these evaluations.

Method	LPIPS	FID
VITON-HD [1]	0.117	12.117
HR-VITON [15]	0.1045	11.265
LADI-VTON [9]	0.0964	9.480
Paint-by-Example [27]	0.1428	11.939
DCI-VTON [4]	0.0804	8.754
GP-VTON [26]	0.088	9.072
Ours	0.0842	8.703
Ours (RePaint [30])	0.0762	8.433

TABLE I: Comparison of LPIPS and FID scores for various virtual try-on methods, including our proposed approach. Lower scores indicate better performance.

In the unpaired context, the FID [28] metric was used to assess the outputs' realism by measuring the statistical similarity between the generated and real picture distributions. With a lower FID score than any of the other models—VITON-HD [1], HR-VITON [15], GP-VTON [26], LADI-VTON [9], and DCI-VTON [4]—EfficientVITON showed exceptional performance in this area. The low FID score shows that our model produces very realistic images that closely match the distribution of real images in the ground truth.

In the paired instance, we evaluated the generated outputs' visual similarity to the ground truth images using the LPIPS [29] metric. In terms of preserving high perceptual quality, EfficientVITON outperformed baseline models and demonstrated competitive performance. However, it is important to note that possible reconstruction mistakes related to the agnostic map utilized in our pipeline caused a slight decline in performance in the paired configuration. To overcome this difficulty, we modified techniques based on RePaint [30] and other methods that maximize the sampling of agnostic zones throughout the inference procedure. The model's ability to preserve areas unrelated to the clothes was greatly improved by these modifications, which also increased the outputs' consistency and LPIPS scores.

Additionally, the results demonstrate how well EfficientVITON generalizes across a variety of experimental setups. Our model consistently produced reliable results, even while rival GAN-based and diffusion-based techniques, such GP-VTON [26] and DCI-VTON [4], displayed performance decreases in specific conditions. This demonstrates how well the architecture of EfficientVITON maintains excellent performance across a range of evaluation methodologies.

In both paired and unpaired settings, the quantitative evaluation clearly shows EfficientVITON's superiority over cutting-edge techniques. The model's capacity to produce visually realistic and perceptually high-quality images is highlighted by its low FID scores and competitive LPIPS performance, which makes it an excellent choice for virtual try-on applications.

4.3 Efficiency Results

In order to assess the efficiency gains achieved by using the non-uniform timestep distribution in our diffusion-based architecture, we thoroughly compared the inference and training times before and after the optimization. (Table II) summarizes the results, which show a significant decrease in computation time, demonstrating the usefulness and efficiency of our technique.

Method	Training time	Inference Time
Before	1570 h	58 s
After	859 h	16 s

TABLE II: Comparison of Training time and Inference time for our model before and After using our new approach of applying the non-uniform timestep distribution.

Before using the non-uniform timestep distribution, the model's inference time for each image was 58 seconds, which reflected the denoising process's computing load. However, the inference time was significantly reduced to just 16 seconds after implementing the optimal timestep sampling approach. With a 72.4% decrease in inference time, this shows that our optimized method effectively boosts the model's performance without sacrificing the caliber of the outputs it produces.

In the same manner, the training procedure experienced a notable increase in effectiveness. The initial diffusion process was resource-intensive, as evidenced by the 1570 hours the model took to train before optimization. The training time was lowered by 45.3% to 859 hours by using the non-uniform timestep distribution. These enhancements greatly increase the training process's usefulness and effectiveness, especially for cases that call for iterative fine-tuning or large datasets.

The results shown in (Table II) clearly demonstrate how well our technique works to lower computational overhead without

sacrificing the quality of the outputs that are produced. The substantial decreases in training and inference times confirm that our approach is feasible for use in practical settings. Furthermore, these developments improve the accessibility and scalability of diffusion-based virtual try-on models by opening the door for their use in settings with constrained computational resources.

EfficientVITON preserves its great performance in both visual appearance and quantitative evaluation, as described in previous sections, while also exhibiting superior efficiency by including the non-uniform timestep distribution. The importance of our contributions to the field of diffusion-based generative models is demonstrated by this comprehensive improvement.

5 Discussion

In this thesis, we developed an advanced virtual try-on system powered by diffusion models, integrating a novel modification that employs non-uniform timesteps in the forward process. This approach was designed to address a critical challenge: reducing computational demands without sacrificing the quality of the generated images. The primary objective of this innovation was to create a system that is computationally efficient while still producing high-fidelity outputs suitable for practical applications.

Our results demonstrate that this modified diffusion model successfully achieved the desired balance. By utilizing non-uniform timesteps, the sampling process became significantly more efficient, requiring far fewer iterations during the forward process. This adjustment directly translated into a substantial reduction in both computational load and training time, enhancing the overall efficiency of the system. Despite these optimizations, the quality of the generated images remained consistently high. This was validated through both qualitative assessments and quantitative metrics, including Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) as shown in Table I.

The qualitative evaluation revealed that the virtual try-on images produced by our modified model were visually indistinguishable from those generated by the standard diffusion model as shown in Table I. Test users reported high levels of satisfaction with the realism, detail, and overall quality of the try-on images, further confirming that the modification did not compromise perceptual quality. On the quantitative side, our FID scores were comparable to the baseline diffusion model, indicating that the use of non-uniform timesteps did not introduce any noticeable artifacts or distortions. Similarly, LPIPS scores verified that the structural integrity of the generated images was preserved, reinforcing the effectiveness of our approach.

Our findings align with and extend prior research on diffusion models and virtual try-on systems. Traditional diffusion models, such as those introduced by Ho et al. [31], rely on uniform timesteps in their forward processes, which often results in high computational costs. More recent work by Song et al. [32] explored the potential of non-uniform timesteps within denoising diffusion probabilistic models (DDPMs), demonstrating that such modifications can expedite the sampling process without sacrificing output quality. Our study builds on this foundation, confirming the advantages of non-uniform timesteps and validating their application in the context of virtual try-on systems.

Additionally, our work contributes to the growing body of research on virtual try-on technologies. Previous approaches, particularly those utilizing Generative Adversarial Networks (GANs), have shown promise but often struggle with maintaining high image fidelity and realism. Diffusion models, on the other hand, are well-suited for generating high-quality images and have emerged as a compelling alternative. Our modification enhances the computational efficiency of diffusion models, addressing one of their main limitations while preserving the high-quality outputs that make them ideal for virtual try-on applications.

The implications of our findings are profound, spanning both academic research and practical applications within the fashion industry. From a research perspective, our study demonstrates that incorporating non-uniform timesteps into diffusion models can effectively reduce computational demands. This opens new avenues for deploying these models in real-time or resource-constrained environments, where efficiency is a critical factor.

For the fashion industry, our virtual try-on system offers a transformative solution for enhancing the online shopping experience. By reducing computational requirements, our model can be more easily integrated into e-commerce platforms, enabling customers to virtually try on clothing with minimal latency. This innovation has the potential to significantly improve customer satisfaction by allowing users to visualize how garments will look and fit before making a purchase.

Moreover, it can lead to higher conversion rates and reduced return rates, as customers gain greater confidence in their buying decisions.

In summary, our work underscores the potential of combining advanced diffusion models with practical modifications, such as non-uniform timesteps, to create systems that are both efficient and high-performing. These findings not only advance the state of research in diffusion models but also pave the way for impactful applications in industries like fashion, where quality and efficiency are equally important.

6 Conclusion

In this work, we introduced EfficientVITON; a cutting-edge virtual try-on system that marks a substantial breakthrough in e-commerce and image synthesis technology. Through the use of diffusion models and innovative adjustments, as the application of non-uniform time steps in the denoising procedure, we were able to effectively tackle two persistent issues in virtual try-on systems: preserving high-quality image production and guaranteeing computational effectiveness. Because of these contributions, EfficientVITON is not only a strong research tool but also an effective option for real applications in settings with limited resources.

Our study showed that conventional diffusion models, although they are effective at producing high-fidelity images, are naturally constrained by their processing requirements because of their uniform timestep distribution. We added non-uniform timestep sampling to the forward process, motivated by recent advancements in diffusion-based generative models. Without impacting the quality of the output images, this optimisation significantly reduces the number of iterations needed. The model can focus on the most important stages of the diffusion process due to the improved sampling approach, which guarantees both effectiveness and visual realism. As a result, EfficientVITON preserves and sometimes even improves the fidelity and perceptual quality of the generated images while achieving a significant reduction in training and inference times (45.3% and 72.4%, respectively).

To further improve the realism and alignment of clothing with the target human body, we included zero cross-attention blocks and a spatial. Without the use of external warping networks, the zero cross-attention blocks provide accurate semantic correlation between human traits and clothing inside the latent space. In the meantime, the spatial encoder makes sure that fine features like fabric qualities, textures, and patterns are preserved. Even for clothing with intricate patterns or unusual structures, this combination enables EfficientVITON to provide aesthetically beautiful and highly accurate virtual try-on images.

We verified EfficientVITON's efficacy against top GAN-based and diffusion-based techniques through comprehensive studies. Our method outperformed previous models, yielding outcomes with lower FID and LPIPS scores, according to quantitative evaluations. These results were further supported by qualitative evaluations, which showed that EfficientVITON regularly produced images that were identical to actual photographs. The alignment, detail preservation, and capacity to accurately depict a variety of human poses, skin tones, and clothing styles without the introduction of artifacts were the main features that were commended for the virtual try-on images.

This work has numerous implications. EfficientVITON is evidence of the ability of optimised diffusion models to strike a compromise between efficiency and quality from a research perspective. The effective use of non-uniform time steps creates new opportunities for speeding up diffusion-based techniques in a variety of applications, such as style transfer, inpainting, and imagine manipulation. We offer a paradigm that future researchers can use to further improve generative models by showing how computational constraints can be resolved without compromising quality.

EfficientVITON provides the fashion and e-commerce sectors with a realistic and expandable answer to one of their biggest problems: developing customized and interesting online shopping experiences. EfficientVITON-powered virtual try-on systems can improve consumer satisfaction by lowering uncertainty in online purchases by enabling users to see clothing on their own photos. This feature could reduce return rates, encourage sustainability by reducing waste and overproduction, and eventually increase e-commerce platforms' profitability. Additionally, because of its efficiency, the system may be deployed on a variety of platforms, including mobile devices and high-end servers, increasing its usefulness and accessibility.

EfficientVITON bridges the gap between high-quality image synthesis and computational efficiency, marking a significant advancement in virtual try-on technology. Its cutting-edge design and remarkable outcomes highlight its potential to raise the bar for virtual try-on systems as a research standard

and a useful tool for real-world applications. EfficientVITON is a lighthouse for future innovation, opening the door for more approachable, effective, and significant solutions in this fascinating topic as the need for sophisticated and effective generative models increases.

7 Future Work

EfficientVITON demonstrates promising results in the virtual try-on domain, but pushing the frontiers of its performance and further increasing its capabilities present compelling areas of future research. A number of lines of research stand out, some very promising as: refining the model architecture to be more complicated, bringing in additional data modalities for a richer virtual try-on experience, and mitigating some existing limitations in complicated scene rendering, accessories, and body shape variables.

This section dives deeper in these promising directions, posing concrete research questions, and giving some ideas for bringing virtual try-on techniques to the next level. These potential improvements promise increased realism of the generated images while increasing the applicability and accessibility of the technology.

7.1 Enhancing Realism and Details

- **Higher Resolution Generation:**
  EfficientVITON already worked well on moderate resolution, however, yet pushing for higher resolutions such as 2048x1536 or higher would further improve the quality of the images generated, enabling better capture of the finer details of clothing, including its textures, patterns, creases, and even accessories. This may be accomplished with further improvements to the model's architecture, training methods or simply better hardware. Also, experimenting with tools such as progressive growing or multi-resolution training may prove useful in this case as well.
- **Improved Fabric Simulation:**
  Simulating the drape and deformation of fabrics for different materials is still complex. Applying physics-based or data-driven methods for more accurate fabric simulation would improve the quality of virtual try on results considerably. These attempts may involve adding the fabric's physical parameters to the model, or adding training data of a fabric's module draped on different body shapes. Working on these attempts can greatly improve the the fabric simulation in EfficientVITON.
- **Handling Complex Clothing Structures:**
  Even with lots of complex structures of clothing, EfficientVITON performs well. However, EfficientVITON's ability is challenged when trying to incorporate more elaborate items of clothing like intricate dresses, flowing dresses, and accessories including belts or scarves. Incorporating specialized attention mechanisms or creating tailored modules within the model to support these structures would provide more flexibility to the system.

7.2 Expanding Functionality and Applications

- **Video-Based Virtual Try-On:**
  Introducing EfficientVITON for real-time video inputs would bring dynamic virtual try-on experiences. Considerations in this respect include maintaining temporal consistency so clothing will realistically drape and align with the person's movements in the video. Recurrent neural networks or temporal attention mechanisms can be employed for the purpose of coherency within the frames.
- **Multi-Modal Input:**
  Moreover, merged modalities like text descriptions or user sketches could further elevate user experience and provide more personalized virtual try-on. The user can describe the desired attributes of the outfit or sketch the outfit, and the model generates virtual try-on images accordingly. Such an approach would require multi-modal encoders and fusion mechanisms in the model architecture.
- **Personalized Recommendations:**
  Using a user's preferences and purchase history data could add a whole new level of value for designing these virtual try-ons. Therefore, these could either include collaborative filtering techniques or could develop user-specific models that'd learn the individual preferences in style.
- **Integration with Augmented Reality (AR):**
  EfficientVITON's implementation with AR technologies would enable users to do a virtual try-on of clothes in real-time using their mobile devices or AR headsets. This would create an immersive and engaging shopping experience, allowing users to visualize how clothing would look on them within their own environment.

7.3 Improving Efficiency and Scalability

- **Further Optimization of the Diffusion Process:** Exploring alternative sampling strategies or noise schedules for the diffusion process may yield some reduction in resource demands and improvements in image quality. This means exploring methods such as importance sampling or constructing more efficient noise schedules especially for the virtual try-on task.
- **Model Compression and Quantization:** Employing model compression techniques like pruning or quantization could result in a smaller model and reduced resources for computations, which makes it suitable for deployment on resource-constrained devices like cell phones.
- **Distributed Training and Inference:** Scaling up the training and inference process using distributed computing frameworks would enable the model to handle even larger datasets and generate higher resolution images more efficiently.

7.4 Addressing Limitations

- **Handling Occlusions and Accessories:** Occluders are the objects or elements in a scene that block or obscure part of another object. In the case of EfficientVITON, examples of Occluders are:
  1) Hair: Long hair falling over the shoulders can block parts of a shirt or dress.
  2) Accessories: Items like necklaces, bracelets, or handbags may overlap with clothing or other body parts.
  3) Body parts: Certain poses (e.g., crossing arms) can obscure parts of the clothing being modeled.

  EfficientVITON may sometimes suffer occlusion and accessory problems that cover parts of clothing, such as hair falling down on the clothing or jewelry worn by a person. Strategies for addressing occluders and for integrating accessories in a more realistic manner would boost the overall robustness for the model. It can include the use of incorporating a depth cue or designing specialized attentional mechanisms focusing on the interactions between clothing, accessory, and body parts.
- **Diverse Body Types and Poses:** While EfficientVITON generalizes fairly well across bodies of various shapes and poses, it deserves further improvements that generalize better to a broader set of body types and challenging poses themselves. These improvements could take the form of model training on more diverse datasets or the formulation of other adaptive schemes in which the model parameters would be adjusted somewhat depending on the properties of the input person.

With these future research directions in mind, EfficientVITON can evolve into a more powerful and versatile tool for virtual try-on to provide an immersive experience and accurate performance by blending digital-with-physical fashion. As these improvements give the system more capabilities to handle complex scenarios such as occlusion, accessory integration, and diverse body types, EfficientVITON has the potential to change how consumers interact with online retail platforms. Essentially, by allowing customers to visualize how garments and accessories will fit and interact with their unique features in real time, it could diminish uncertainties and promote sound decision-making during online shopping. Such improvements would eventually lead to greater consumer satisfaction while also reducing return rates, thus offering a sustainable approach to fashion retail. EfficientVITON can change the very nature of the online shopping experience and set new benchmarks for convenience, personalization, and innovation within the fashion industry.

Acknowledgment

The authors would like to express their gratitude to IONO Tech Company for their invaluable support and a very cooperative sponsorship. IONO Tech not only supplied this research with the necessary hardware resources, but also offered helpful and informative guidance throughout the process.

References

[1] S. Choi, S. Park, M. Lee, and J. Choo, "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 131–14 140.

[2] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: an open large-scale dataset for training next generation image-text models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2024.

[3] 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.

[4] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, "Taming the power of diffusion models for high-quality virtual try-on with appearance flow," *arXiv preprint arXiv:2308.06101*, 2023.

[5] 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 (CVPR)*, 2021, pp. 8485–8493.

[6] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 204–219.

[7] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 589–604.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[9] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on," *arXiv preprint arXiv:2305.13501*, 2023.

[10] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[11] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman, "Tryondiffusion: A tale of two unets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4606–4615.

[12] X. Han, Z. Wu, Z. Wu *et al.*, "Viton: An image-based virtual try-on network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7543–7552.

[13] B. Wang, H. Zheng, X. Liang *et al.*, "Toward characteristic-preserving image-based virtual try-on network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 589–604.

[14] T. Issenhuth, J. Mary, and C. Calauzenes, "Do not mask what you do not need to mask: A parser-free virtual try-on," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 619–635.

[15] Q. Lyu, Q. Wang, and K. Huang, "High-resolution virtual try-on network with coarse-to-fine strategy," *Journal of Physics: Conference Series*, vol. 1880, p. 012009, 04 2021.

[16] B. Ren, H. Tang, F. Meng *et al.*, "Cloth interactive transformer for virtual try-on," *arXiv preprint arXiv:2104.05519*, 2021.

[17] Z. Xie, Z. Huang, F. Zhao *et al.*, "Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 2598–2610.

[18] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[20] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[22] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[23] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6757–6765.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335280

[25] H. Jiang, M. Imran, L. Ma, T. Zhang, Y. Zhou, M. Liang, K. Gong, and W. Shao, "Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation," *arXiv preprint arXiv:2405.14802*, 2024.

[26] 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," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23 550–23 559, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257757040

[27] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 381–18 391, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253802085

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *ArXiv*, vol. abs/1706.08500, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:231697514

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4766599

[30] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 451–11 461, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246240274

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arxiv:2006.11239*, 2020.

[32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

APPENDIX A
FURTHER DISCUSSION OF EFFICIENTVITON OUTPUT

In this appendix, EfficientVITON's output is further discussed and explored. In Fig. 22, a table of different combinations of persons and garments is shown. It can be observed that EfficientVITON's output is robust to different person poses and skin tones, due to the model's use of the person's DensePose as input. Additionally, it can be noticed that the model is able to generate images of highly textured clothes without any problem or loss of data, as the Spatial Encoder helps preserve garment details during the denoising process.

Fig. 22: Person-Cloth Combinations. Best viewed when zoomed in.