clothes

EfficientVITON

An Efficient Virtual Try-On Model using Optimized Diffusion Process

Advisor : Prof. Ahmed Fares

Date : 05/02/2025

E-JUST
Egypt - Japan University of Science and Technology
الجامعة المصرية اليابانية للعلوم والتكنولوجيا
エジプト日本科学技術大学

IONO TECHNOLOGY

Meet the Team

Ashrakat Saeed
120200091

Abdelrahman Said
120200075

Ahmed Rashed
120200225

Mariam Ayman
120200094

Mostafa Atef
120200043

Sponsored by

IONO TECHNOLOGY

Agenda

- Introduction
- Related Work
- Methodology
- Results & Discussion
- Conclusion & Future Work

Agenda

- Introduction
- Related Work
- Methodology
- Results & Discussion
- Conclusion & Future Work

Introduction

Motivation

- Reduce e-commerce return rates (**up to 30% due to try-on issues).**
- Promote sustainability by minimizing physical try-ons.

Problem Statment

Traditional virtual try-on methods heavily **depend on paired datasets and external warping networks** to align garments with the human body. However, these approaches face several critical challenges that limit their effectiveness and scalability.

Introduction

Virtual try-on systems face three key challenges

Challenge 1:

Detail Preservation

Maintaining garment details such as textures, patterns, and fabric folds remains a challenge, especially in high-resolution applications.

Challenge 2:

Computational Efficiency

The iterative nature of many virtual try-on methods results in high computational demands, making real-time performance impractical.

Challenge 3:

Realistic Alignment

Establishing precise correspondence between garments and body features.

Agenda

- Introduction
- Related Work
- Methodology
- Results & Discussion
- Conclusion & Future Work

Related Work

Evolution of Virtual Try-On Methods

Geometric Warping (e.g., VITON)

✅ Low computational cost.

❌ Poor detail preservation.

GAN-Based Models (e.g., CP-VTON)

✅ Realistic outputs.

❌ Artifacts and unstable training.

Diffusion Models (e.g., TryOnDiffusion):

✅ High-fidelity details.

❌ Slow inference (e.g., 58s/image).

Research Gaps

- Trade-off between speed and quality.
- Limited generalization to diverse poses/body types.

Related Work

Model	Approach	Advantages	Limitations	Efficiency
CP-VTON	GAN-based, Warping	Generates realistic images, handles complex clothing patterns	Computationally expensive, can be unstable during training, may struggle with diverse poses	Moderate
VITON-HD	Warping-based	Relatively fast, good for simple clothing items	Limited ability to handle complex clothing textures/patterns, struggles with occlusions and challenging poses	High
TryOnDiffusion	Diffusion-based	High realism, handles complex scenes and poses	Requires large paired datasets, computationally expensive	Low
DCI-VTON	Diffusion-based, Warping	Uses pre-trained diffusion model, good realism	Relies on external warping network, inherits limitations of warping approaches	Moderate

Agenda

- Introduction
- Related Work
- Methodology
- Results & Discussion
- Conclusion & Future Work

clothes

Methodology

01

Data Preprocessing

02

EfficientVITON Model Architecture

03

Optimized Diffusion Process

04

Spatial Encoder and Zero Cross-Attention Blocks

05

Loss Functions and Training

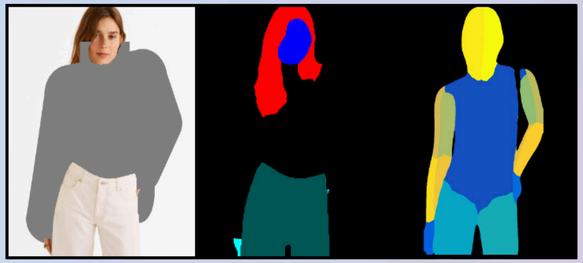
IONO TECHNOLOGY

clothes

Data Preprocessing

Data Preprocessing

a b c



d e f

g h

01

The main training dataset used was VITON-HD. This dataset was a contribution for the research in one of the first virtual try-on researches by Choi et al.

02

The dataset contains over **11600 training images** and over **2000 testing images.**

03

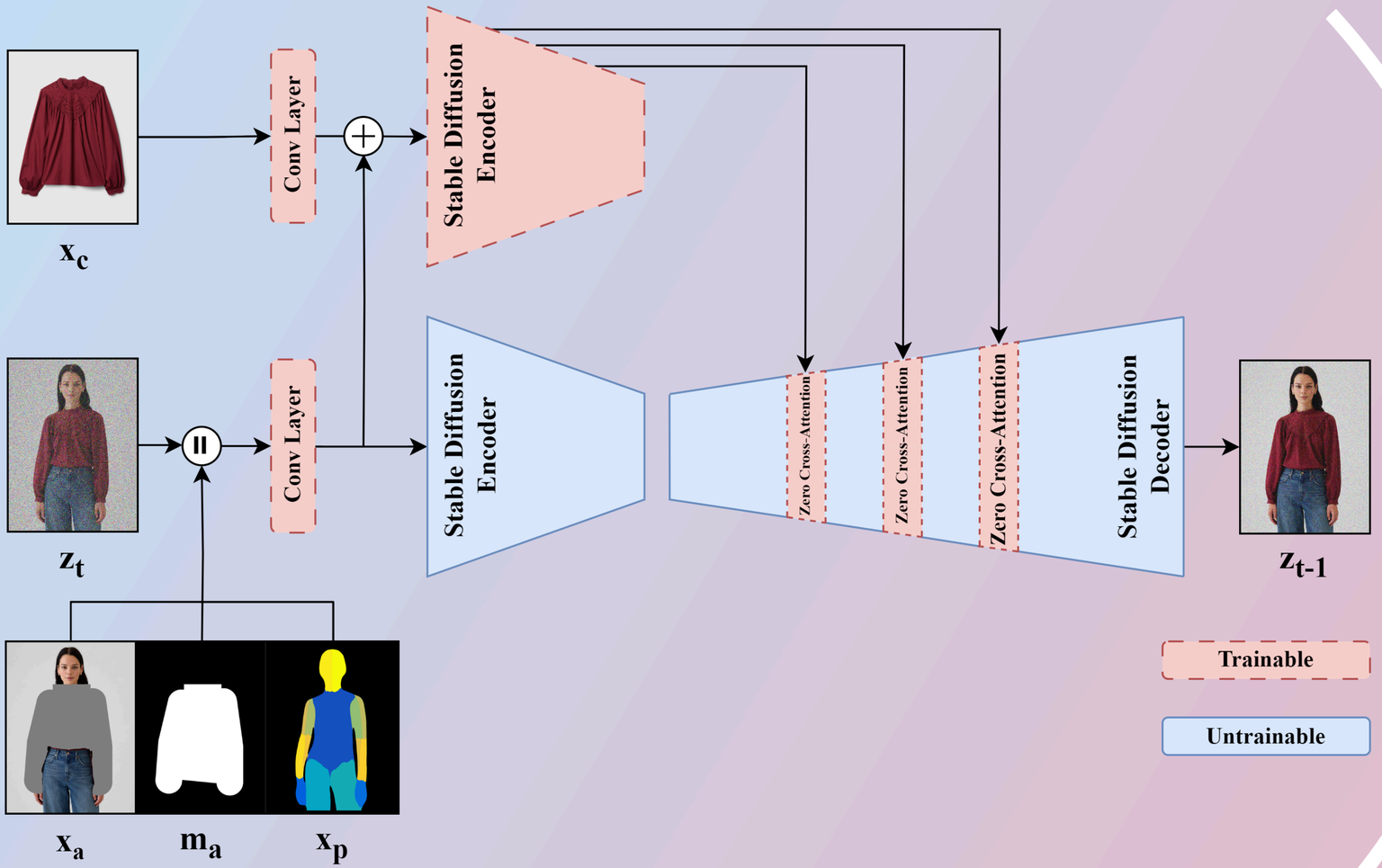
All images **are 1024x768** in terms of resolution. Each image in the dataset has its equivalent image of the unworn clothes.

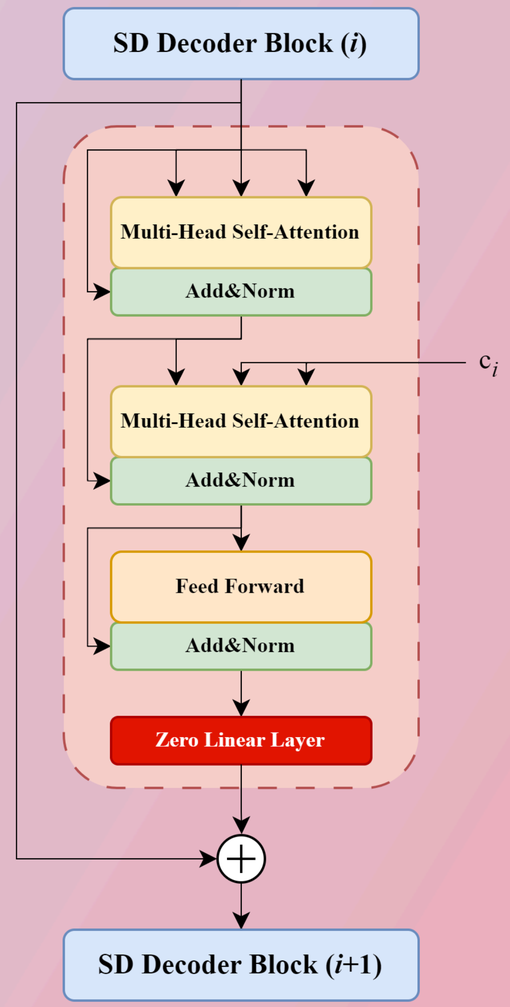
04

In addition, the dataset handles most of the preprocessing steps, like the densepose, the human parsing, the pose estimation, etc.

clothes

EfficientVITON Model Architecture





IONO TECHNOLOGY

Optimized Diffusion Process

- **Base Model:** Pre-trained Stable Diffusion v1.5 - Foundation for high-quality image generation.
- **Sampling Strategy:** Improved DDPM Sampler - Reduced sampling steps for faster inference while maintaining quality. Instead of 50 uniformly distributed timesteps, we apply a non-uniform distribution of 10 timesteps in the denoising process.

	LPIPS	FID
Uniform Timesteps (50)	0.0732	8.233
Non-uniform Timesteps (10)	0.0762	8.433

- **Latent Space Optimization:** Direct manipulation in latent space - Efficient transformations and warping for clothing.

clothes

Spatial Encoder & Zero Cross-Attention Blocks

الجامعة المصرية اليابانية للعلوم والتكنولوجيا
E-JUST
Egypt - Japan University of Science and Technology
エジプト日本科学技術大学

Spatial Encoder

- The spatial encoder is initialized with the U-Net Encoder weights of the pretrained Stable Diffusion Model.
- It produces feature maps at multiple resolutions as it downsamples the input clothing image.

Zero Cross Attention Block

- Learns the relationship, or semantic correspondence, between the features of the clothing (extracted by the spatial encoder) and the regions of the person's body (represented by the U-Net's feature maps).
- Patch-wise Warping: Transforms and warps clothing in latent space for precise fitting

Loss Functions & Training

- **Latent Diffusion Loss (LLDM)** - Guides the denoising process in the latent space.

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

Where:

- **ζ:** Concatenated latent inputs (noisy latent map zt , latent agnostic map E(xa), agnostic mask xma , latent dense pose E(xp)).
- **xc :** Clothing image.
- **ϵ:** Random noise.
- **ϵθ :** U-Net parameterized by θ.
- **t:** Timestep.
- **Tϕ :** CLIP image encoder parameterized by ϕ.

Loss Functions & Training

- **Attention Total Variation Loss (LATV)** - Sharpens attention, enhances clothing detail.

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

Where:

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

- **Overall Loss:** $L_{total} = L_{LDM} + \lambda_{ATV} L_{ATV}$

Where:

- λATV is a hyperparameter weighting the LATV

Loss Functions & Training

EfficientVITON's training process likely involves **two** stages:

1

Semantic Correspondence Learning

- The model focuses on learning the semantic relationship between clothing and body parts.
- The LLDM is the primary loss function during this stage.

2

Attention Refinement

- The LATV loss is introduced to refine the attention maps.
- This helps to sharpen the attention and improve the precision of clothing placement.

clothes

model more robust, like Random Scaling, Random Rotation, Color Adjustments.

Agenda

- Introduction
- Related Work
- Methodology
- Discussion and Results
- Conclusion & Future Work

Results & Discussion (Quantitative)

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

EfficientVITON achieves state-of-the-art FID scores and **cuts inference time to 16 seconds**, making real-time applications feasible.

clothes

Results & Discussion (Quantitative)

	LPIPS	FID
Uniform Timesteps (50)	0.0732	8.233
Non-uniform Timesteps (10)	0.0762	8.433

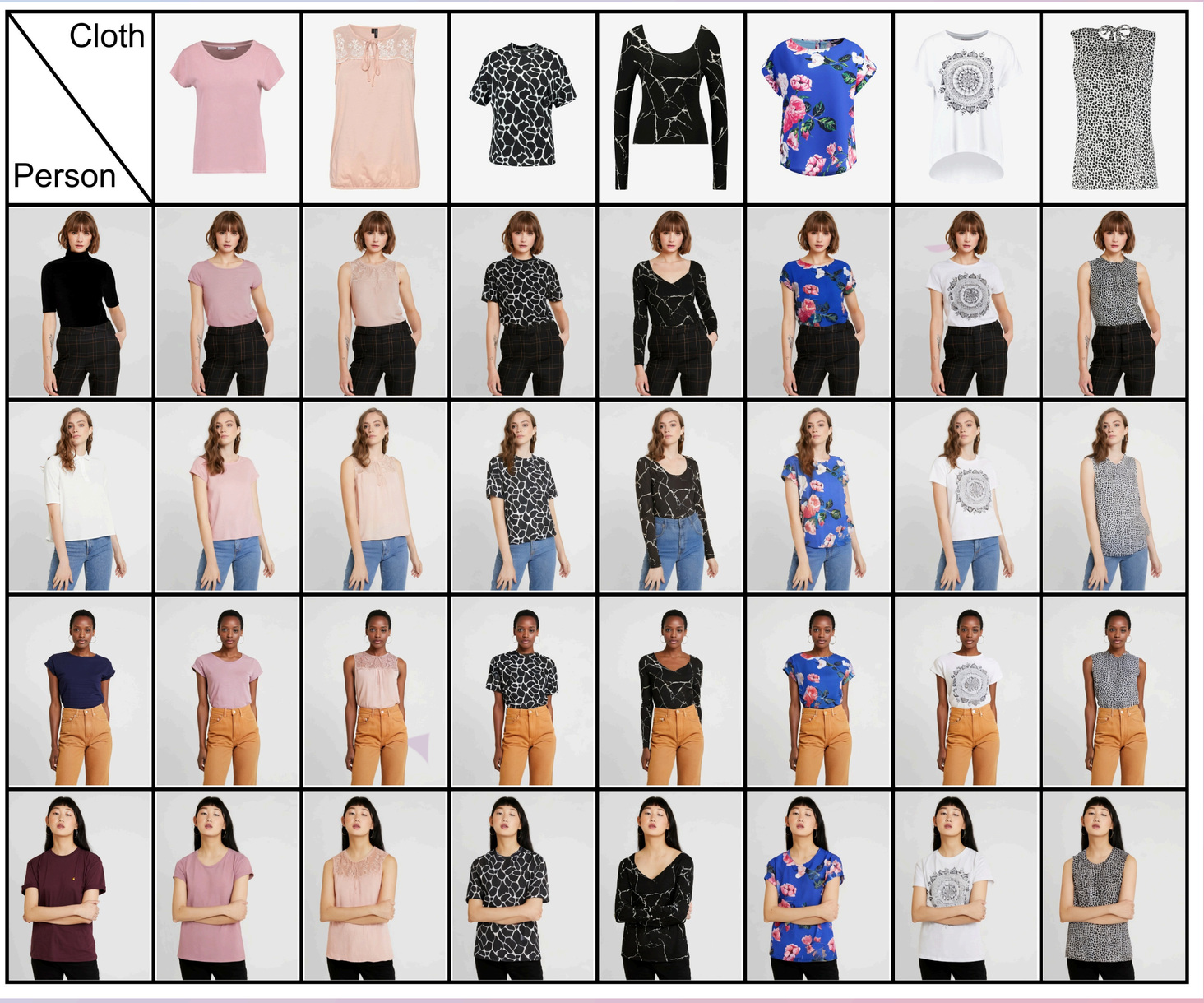
Method	Training	Inference
Before	1570 h	58 s
After	859 h	16 s

clothes

Results & Discussion (Qualitative)

- Accurate alignment **(no warping)**.
- Preserved textures **(logos, patterns)**.

clothes



الجامعة المصرية اليابانية للعلوم والتكنولوجيا
E-JUST
Egypt - Japan University of Science and Technology
エジプト日本科学技術大学

IONO TECHNOLOGY

Agenda

- Introduction
- Related Work
- Methodology
- Results & Discussion
- Conclusion & Future Work

clothes

Conclusion

Efficiency and Accuracy

Improves Efficiency

EfficientVITON significantly improves the efficiency of virtual try-on while maintaining accuracy. Working within the compressed latent space further reduces computational overhead and allows for semantically meaningful clothing manipulations.

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.

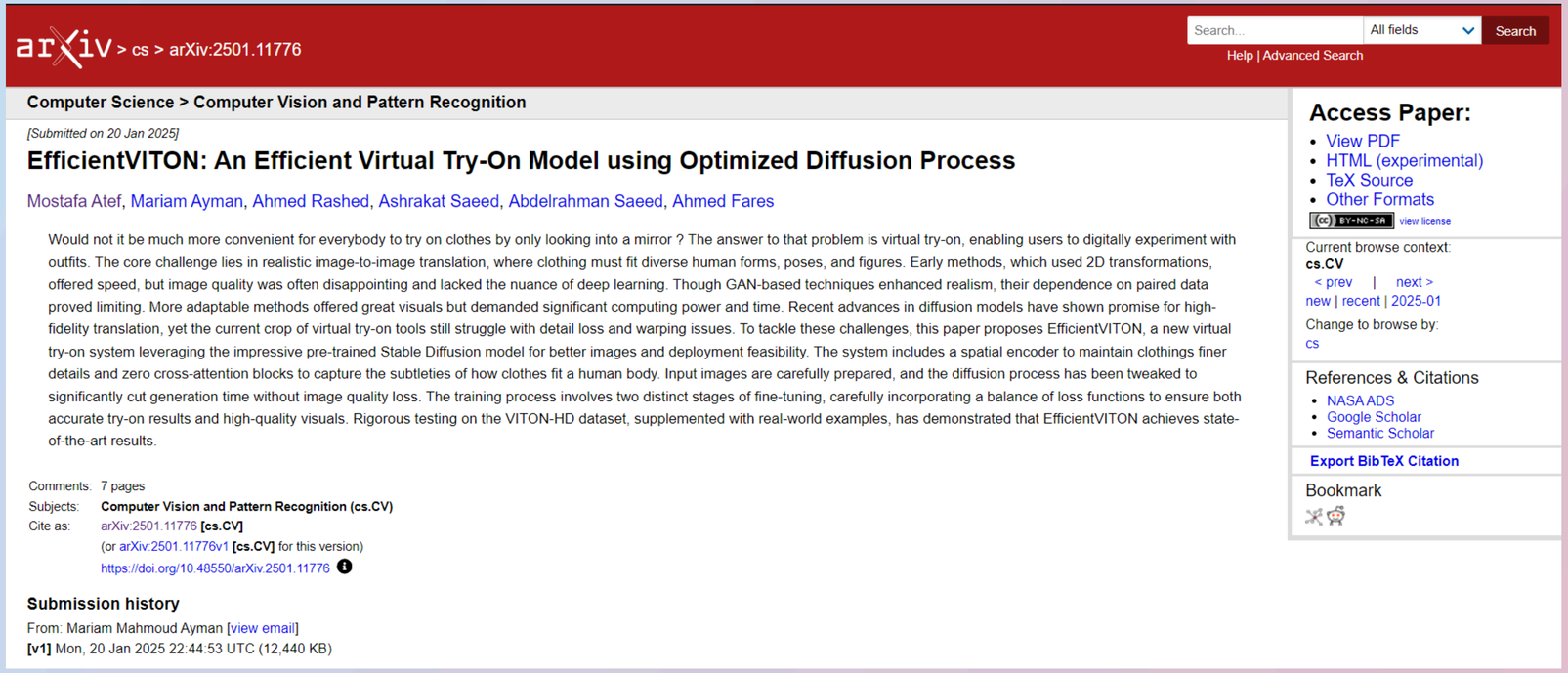
Detailed Clothing Preservation

EfficientVITON excels at preserving intricate clothing details. The spatial encoder, initialized with U-Net weights, effectively captures fine-grained features such as textures, patterns, and logos.

The zero cross-attention blocks, learning the precise correspondence between clothing and body parts, enable highly accurate and realistic warping of clothing onto the person, resulting in a natural and convincing try-on effect.

Conclusion

Impact and Publication



arXiv > cs > arXiv:2501.11776

Search... All fields Search

Help | Advanced Search

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 20 Jan 2025]

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

Mostafa Atef, Mariam Ayman, Ahmed Rashed, Ashrakat Saeed, Abdelrahman Saeed, Ahmed Fares

Would not it be much more convenient for everybody 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, which used 2D transformations, offered 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 paper proposes EfficientVITON, a new virtual try-on system leveraging the impressive pre-trained Stable Diffusion model for better images and deployment feasibility. The system includes a spatial encoder to maintain clothings 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.

Comments: 7 pages
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**
Cite as: arXiv:2501.11776 **[cs.CV]**
(or arXiv:2501.11776v1 **[cs.CV]** for this version)
https://doi.org/10.48550/arXiv.2501.11776

Submission history
From: Mariam Mahmoud Ayman [view email]
[v1] Mon, 20 Jan 2025 22:44:53 UTC (12,440 KB)

Access Paper:
- View PDF
- HTML (experimental)
- TeX Source
- Other Formats

CC BY-NC-SA view license

Current browse context:
cs.CV
< prev | next >
new | recent | 2025-01

Change to browse by:
cs

References & Citations
- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark

We believe **EfficientVITON makes a significant contribution to the field of virtual try-on.** Our work has been published on arXiv and we are currently preparing a manuscript for submission to a leading conference/journal in computer vision.

Future Work

Working on Dynamic Garment Simulation

To enhance realism further, we plan to incorporate physics-based simulations to model the dynamic behavior of clothing. This will allow EfficientVITON to capture the natural drape, folds, and wrinkles of garments, leading to even more convincing virtual try-on experiences.

Expanding on Diverse Datasets

Training on more diverse and representative datasets is crucial. We aim to expand our training data to include a wider range of body types, ethnicities, and clothing styles, thereby improving the generalizability and inclusivity of EfficientVITON.

clothes

E-JUST
Egypt - Japan University of Science and Technology

Thank You

-ANY QUESTIONS?-

IONO TECHNOLOGY