DIFFUSION BASED VIRTUAL TRY ON: DiMVTON

R.S.N. de Zoysa

228851T

Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2025

DIFFUSION BASED VIRTUAL TRY ON: DiMVTON

R.S.N. de Zoysa

228851T

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree

Master of Science in Artificial Intelligence

Department of Computational Mathematics

Faculty of Information Technology

University of Moratuwa

Sri Lanka

April 2025

DECLARATION

I hereby declare that this thesis/dissertation is the result of my own work and does not include, without proper acknowledgment, any material previously submitted for a degree or diploma at any other university or institution of higher education. To the best of my knowledge, it does not contain any material previously published or authored by another individual, except where due acknowledgment is given within the text. I reserve the right to reuse this work, in whole or in part, in future publications such as articles or books.

Signature: Date: 01/07/2025

The candidate named above has conducted the research presented in this master's thesis/dissertation under my supervision. I hereby confirm that the declaration provided by the student is accurate and truthful.

Name of Supervisor: Prof A T P Silva

Signature of the Supervisor: Date: 2/07/2025

DEDICATION

To my mother and family, for instilling in me the values of perseverance and curiosity.

To my husband, for his unwavering encouragement, belief in my dreams, and constant motivation through every challenge.

To my mentors and teachers, for inspiring me to pursue knowledge and innovation.

To all researchers and engineers dedicated to advancing virtual try-on technology, I hope this work serves as a contribution toward the development of more efficient, scalable, and accessible AI-powered solutions.

ACKNOWLEDGEMENT

I would like to sincerely thank my advisor, Professor ATP Silva, for her exceptional guidance, support, and encouragement throughout the course of this research. Her expertise and thoughtful insights have been pivotal in shaping the direction and quality of this work.

I also wish to extend my gratitude to the creators and maintainers of the VITON-HD public dataset, whose efforts have been fundamental for benchmarking and evaluating the proposed method.

In addition, I acknowledge the developers of open-source machine learning frameworks, whose tools and resources made this research possible.

Finally, I am profoundly grateful to my family, and friends for their unwavering support, patience, and belief in me throughout this journey. My husband's constant encouragement, understanding, and steadfast support have been instrumental in helping me remain focused and motivated during the challenges of this research.

ABSTRACT

Diffusion models have recently set new standards for realism in virtual try-on tasks, yet most existing systems are burdened by the need for additional modules such as Reference Networks, complex image/text encoders, and heavy preprocessing pipelines. These extra components substantially increase the number of trainable parameters, GPU memory consumption, and overall computational cost. In this paper, we introduce DiMVTON, a highly efficient diffusion-based framework for virtual try-on that rethinks this complexity. Instead of relying on external conditioning networks, DiMVTON simply concatenates person and garment inputs along the spatial dimension and feeds them directly into a streamlined denoising UNet. Our approach is driven by three main efficiency goals: (1) Compact architecture - DiMVTON uses only a VAE and a minimal UNet without cross-attention or external encoders, achieving a total model size of 894.29 million parameters. (2) Selective fine-tuning - comprehensive studies reveal that the UNet's self-attention layers are the critical elements for aligning garments onto individuals. Fine-tuning only these layers enables strong performance with just 0.39 million trainable parameters (around 0.04% of the backbone), as its further optimized through Low-Rank Adaptation (LoRA) techniques. (3) Minimal inference overhead – Unlike other diffusion-based models that require auxiliary information like human parsing maps, pose annotations, or textual descriptions, DiMVTON needs only a person image, a garment reference, and a simple mask, cutting memory usage by over 90%. Despite being trained on a relatively small dataset of 13,000 samples, DiMVTON achieves competitive qualitative and quantitative results and shows strong generalization in real-world scenarios. Our findings suggest that high-quality virtual try-on is possible without complex architectures, provided that fine-tuning is applied strategically to key network components.

Keywords: Diffusion models, Virtual try-on, Efficient training, Self-attention layers, LoRA fine-tuning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
VTON	Virtual Try-On
LDM	Latent Diffusion Model
VAE	Variational Autoencoder
LoRA	Low-Rank Adaptation
UNet	U-shaped Convolutional Neural Network
DDIM	Denoising Diffusion Implicit Models
SSIM	Structural Similarity Index Measure
LPIPS	Learned Perceptual Image Patch Similarity
FID	Fréchet Inception Distance
KID	Kernel Inception Distance
MSE	Mean Squared Error
CLIP	Contrastive Language–Image Pre-training
DINOv2	Self-supervised ViT Model (from Meta)
GAN	Generative Adversarial Network
TPS	Thin Plate Spline
DREAM	Diffusion Rectification and Estimation-Adaptive Models
OOTDiffusion	Out-of-the-Tailored Diffusion
RePaint	Iterative Inpainting Diffusion Model

LIST OF APPENDICES