

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

DECLARATION.....	i
DEDICATION.....	ii
ACKNOWLEDGEMENT.....	iii
ABSTRACT.....	iv
TABLE OF CONTENTS.....	v
LIST OF FIGURES.....	vi
LIST OF TABLES.....	vii
LIST OF ABBREVIATIONS.....	viii
LIST OF APPENDICES.....	ix
CHAPTER 1.....	1
INTRODUCTION.....	1
1.1 Prolegomena.....	1
1.2 Aim and Objective.....	2
1.3 Background and motivation.....	2
1.4 Problem in brief.....	4
1.5 Structure of the Thesis.....	4
1.6 Summary.....	5
CHAPTER 2.....	7
LITERATURE REVIEW.....	7
2.1 Introduction.....	7
2.2 Subject driven image generation.....	9
2.3 Warping-Based Virtual Try-On Methods.....	10
2.3.1 Thin Plate Splines (TPS)-Based Warping.....	10
2.2.2 Flow-Based Warping.....	11
2.2.3 Comparison of Warping-Based Methods.....	12
2.4 GAN-Based Virtual Try-On Models.....	13
2.4.1 Conditional GANs for Virtual Try-On.....	14
2.4.2 Parsing-Based GANs.....	15

2.3.3 Challenges and Limitations of GAN-Based VTON.....	15
2.3.4 Comparison of GAN-Based VTON Approaches.....	16
2.3.5 Transition to Diffusion-Based VTON.....	16
2.5 Diffusion-Based Virtual Try-On Methods.....	17
2.5.1 Latent Diffusion Models (LDMs) for VTON.....	17
2.5.2 Comparison of Diffusion-Based VTON Approaches.....	19
2.5.4 Challenges and Future Directions for Diffusion-Based VTON.....	20
2.6 Image Inpainting for Virtual Try-On.....	21
2.6.1 Traditional Inpainting Methods.....	21
2.6.2 Diffusion-Based Inpainting.....	21
2.7 Research Gap and Motivation for DiMVTON.....	22
2.7.1 How DiMVTON Addresses These Gaps.....	24
2.8 Summary.....	25
CHAPTER 3.....	27
TECHNOLOGY.....	27
3.1 Introduction.....	27
3.2 Diffusion Probabilistic Models.....	27
3.2.1 Overview of Diffusion Models.....	27
3.2.2 Latent Diffusion Models (LDMs) and Stable Diffusion.....	29
3.2.3 DREAM (Diffusion Rectification and Estimation-Adaptive Models)..	31
3.2.4 Integration in DiMVTON.....	31
3.3 Image Inpainting for Virtual Try-On.....	31
3.3.1 Diffusion-Based Inpainting.....	31
3.3.2 How DiMVTON Uses Diffusion-Based Inpainting.....	32
3.4 LoRA for Parameter-Efficient Fine-Tuning.....	32
3.5 Summary.....	33
CHAPTER 4.....	34
APPROACH.....	34
4.1 Introduction.....	34

4.2 Hypothesis.....	34
4.3 Input.....	35
4.4 Output.....	35
4.5 Process.....	35
4.6 Users.....	36
4.7 Features.....	36
4.8 Summary.....	36
CHAPTER 5.....	37
DESIGN.....	37
5.1 Introduction.....	37
5.2 Architectural Design.....	38
5.3 Data Flow and Process Design.....	38
5.4 Architectural Design Components.....	40
5.4.1 VAE Encoder and Decoder.....	40
5.4.2 Noise Injection in Latent Space.....	40
5.4.3 U-Net Diffusion Process.....	40
5.4.4 LoRA-Based Adaptation.....	41
5.4.5 Training Design.....	41
5.4.6 Loss Functions.....	41
5.5 Design Summary.....	42
CHAPTER 6.....	43
IMPLEMENTATION.....	43
6.1 Introduction.....	43
6.2 Dataset.....	44
6.2.1 Data Set Structure.....	44
6.2.2 Key Features.....	45
6.2.3 Data Set Contents.....	45
6.3 Resource Requirements.....	47
6.3.1 Hardware Requirements.....	47

6.3.2 Software Stack.....	48
6.3.3 Parameters Used.....	48
6.4 Training Pipeline.....	48
6.4.1 Data Loading.....	48
6.4.2 Latent Encoding.....	49
6.4.3 Noise Addition.....	49
6.4.4 Diffusion Denoising.....	49
6.4.5 Reconstruction.....	49
6.4.6 Checkpointing and Logging.....	49
6.5 Computational Efficiency strategy implementation.....	49
6.6 Model Integration.....	50
6.7 Implementation Summary.....	51
CHAPTER 7.....	52
EVALUATION.....	52
7.1 Introduction.....	52
7.2 Qualitative comparison.....	52
7.2.1 Human Diversity Features.....	52
7.2.2 Garment Features.....	53
7.2.3 Qualitative Assessments.....	53
7.2.4 Qualitative Benchmarking against SOTA.....	56
7.3 Evaluated Metrics.....	57
7.3.1 Structural Similarity Index (SSIM).....	57
7.3.2 Fréchet Inception Distance (FID).....	58
7.3.3 Kernel Inception Distance (KID).....	58
7.3.4 Learned Perceptual Image Patch Similarity (LPIPS).....	59
7.4 Quantitative Evaluation.....	59
7.5 Training Evaluation.....	60
7.5.1 Training Hyperparameters.....	60
7.5.2 Training and Testing Loss Behaviour.....	61

7.6 Computational Efficiency.....	62
7.7 Evaluation Summary.....	64
CHAPTER 8.....	65
CONCLUSION.....	65
8.1 Introduction.....	65
8.2 Conclusion.....	65
8.3 Future Work.....	66
REFERENCES.....	68
APPENDICES.....	75

LIST OF FIGURES

Figure	Description	Page
Figure 3.2.1.1	Forward and reverse diffusion overview	28
Figure 3.2.1.2	Step-by-step diffusion process visualization	28
Figure 3.2.1.3	Forward and reverse diffusion pipeline with UNet	29
Figure 5.2	DiMVTON model architecture	38
Figure 7.2.3	Example samples and resulting output of the try-on data	54
Figure 7.2.4	Qualitative comparison against SOTA	56
Figure 7.5.2.1	Training loss curve	61
Figure 7.5.2.2	Validation loss curve	62
Figure 7.5.2.3	Gradient evolution during training	62
Figure A.1	Sample Visual Result 1	75
Figure A.2	Sample Visual Result 2	76
Figure A.3	Sample Visual Result 3	76
Figure A.4	Sample Visual Result 4	77
Figure A.5	Sample Visual Result 5	77

LIST OF TABLES

Table	Description	Page
Table 2.2.3	Comparison of warping-based VTON me	12
Table 2.3.4	Comparison of GAN-based VTON methods.	16
Table 2.5.2	Comparison of diffusion-based VTON methods.	19
Table 4.6	Users and Applications of virtual try-on technology.	36
Table 4.7	Key features of the DiMVTON model.	36
Table 6.2.2	Dataset statistics and properties.	45
Table 6.2.3	Comparison of paired and unpaired training modes.	47
Table 6.5	Optimization strategies.	50
Table 6.6	Model architecture and training settings.	50
Table 6.7	Evaluation metrics and interpretation.	57
Table 7.4	Quantitative comparison with other methods.	59
Table 7.5.1	Training hyperparameters.	60

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

Appendix	Description	Page
Appendix A	Sample Visual Results	75
Appendix B	DiMVTON Model Architecture	80
Appendix C	Virtual Try-On Inference Pipelines	83
Appendix D	LoRA Module for Attention Layers	89
Appendix E	Model-Specific Utility Functions	90
Appendix F	DiMVTON Evaluation Dataset Loader	91
Appendix G	DiMVTON Training Script	93
Appendix H	General Utility Functions	96
Appendix I	DiMVTON Inference Script	101