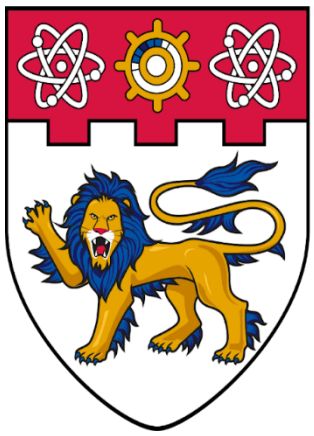


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

COLLEGE OF COMPUTING AND DATA SCIENCE

Final Year Project

CCDS24-0693

Human Pose Transfer via Pre-trained Text-to-Image Models

Peh Jin Yang

Supervisor: Assistant Professor Pan Xingang

Examiner: Assistant Professor Zhao Jun

2025

Abstract

Diffusion model is a promising approach for image generation and applications of the diffusion model for Human Pose Transfer have yielded competitive performance. One such Diffusion model is Stable Diffusion, which is a pre-trained Text-to-Image model that has gained traction for its superior generative capabilities. In recent developments, image adapters such as ControlNet and IP-Adapter have also gained traction for being able to condition the image generation outputs using image prompts while relying only on a small number of additional training parameters.

While most existing works are based purely on image prompts, to the best of this report's knowledge, works built upon Stable Diffusion and image adapters utilising both text and image prompts to perform the task of Human Pose Transfer are limited. Moreover, existing works tend to fail to generalise over a diverse range of samples.

In this report, a novel model architecture that incorporates image adapters with the Stable Diffusion backbone has been proposed. Even though the proposed model achieves slightly sub-par performance as compared to existing baselines, it provides promising opportunities for harmonising text and image prompts for the task of Human Pose Transfer and is also able to achieve decent performance on samples that it was not trained on. This could serve as a stepping stone for future applications in performing Human Pose Transfer using prompts of multi-modality across a diverse range of settings.

Acknowledgements

I would like to offer my sincere gratitude for the opportunity to work on this project as it has allowed me to explore various state-of-the-art models used today in the field of image generation. Moreover, it has allowed me to gain invaluable experience in the field of GenAI, managing and executing a research-focused machine learning project.

Next, I would like to express my sincere gratitude to several individuals for supporting me throughout this project. This project would not have been possible without them.

I would like to express my deepest thanks to Assistant Professor Pan Xingang for providing me with this opportunity to work on this project. His support and guidance have helped me progress in this project.

I would also like to extend my sincerest gratitude to Research Associate Ouyang Wenqi for her patience and guidance throughout the span of the entire project. She readily offers insights, resources, and gives constructive feedback to guide me throughout the project.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

1.1.1 Human Pose Transfer

Over the past decade, significant progress in the image synthesis has been made possible with deep generative models e.g., Generative Adversarial Networks and Variational Adversarial Networks. These developments have sparked great interest in different tasks involving human image generation, each with its own constraints. Examples of some of these constraints include generating humans with diverse clothing attributes or generating humans from different viewpoints, which is also known as Human Pose Transfer.

This project focuses on the task of Human Pose Transfer, which aims to generate images of humans in different poses while maintaining the subject's stylistic attributes. Human Pose Transfer remains an active and developing field of research due to its vast potential applications, such as virtual try-ons (VITON), animated films and games, and virtual human avatar services in the metaverse [1].

Given a human image influencing the pose (i.e., the posture of the subject) and another human image influencing the stylistic attributes (i.e., clothing and facial features of the subject), the aim is to generate a human image incorporating both the pose and styles.

1.1.2 Pre-trained Text-to-Image models

In recent years, large Text-to-Image models like Stable Diffusion [2] have demonstrated their superior capability in image generation. Built on top of CompVis's VAE and OpenAI's text encoder, Stable Diffusion is a diffusion-based generative model trained on the LAION-5B dataset [3]. This dataset consists of 5.85 billion CLIP-filtered web-sourced images matched with associated captions or metadata that describe the visual content.

However, Stable Diffusion may not be proficient in specific tasks such as Human Pose Transfer as it was trained on large-scale generic image-text pairs. Besides reasonable image generation capability, the capability of upholding structural and visual style constraints must be met as well in Human Pose Transfer.

Fortunately, structural conditioning was made possible with ControlNet [4]. By passing in the image prompt containing the structural conditioning (i.e. pose) to a trainable copy of Stable Diffusion U-Net's input blocks followed by zero convolutions, generating human images with the subject in a specific pose is made possible. Around the same time, IP-Adapter [5] proposed a novel decoupled cross-attention mechanism that involves performing additional cross-attention on an image prompt in the Stable Diffusion U-Net's attention layers to incorporate the styles present in the image prompt in the final generated image.

Together, these developments provide an exciting opportunity to leverage the superior image generation capabilities of Stable Diffusion to perform Human Pose Transfer, which requires precise conditioned generation of human images conforming to the pose and style.

1.2 Objectives and Aims

Despite the recent developments in structural and stylistic conditioning made available with image prompt adapters, there still exists a gap in previous works that utilises Stable Diffusion with image prompt adapters (i.e. ControlNet and IP-Adapter) for the task of Human Pose Transfer. While some of the latest developments concerning Human Pose Transfer build upon a pre-trained Stable Diffusion backbone, to the best of this report's knowledge, existing works focusing on the task of Human Pose Transfer using a Stable Diffusion backbone with image prompt adapters are limited.

Hence, the end goal of this project is to develop a novel model architecture incorporating both image prompt adapters and the pre-trained Text-to-Image Stable Diffusion backbone to perform Human Pose Transfer. By effectively integrating the structural conditioning provided by ControlNet and the stylistic conditioning provided by IP-Adapter with the pre-trained Stable Diffusion backbone, it is hoped that the model would be able to generate high-quality human images that conform to the stylistic attributes of a source image and pose of a target image. At the same time, text prompts could be utilised to reinforce stylistic attributes or capture more stylistic attributes without causing entanglement with the image prompts.

To accomplish this, the following outcomes are proposed:

1. Develop a naïve implementation of a pipeline utilising Stable Diffusion with ControlNet and IP-Adapter to perform Human Pose Transfer.
2. Perform a literature review on existing methods used for Human Pose Transfer and shortlist a few state-of-the-art models for comparison with the pipeline built in #1.
3. Improve the pipeline built in #1 to develop a more sophisticated pipeline incorporating the concepts to condition generation outputs using ControlNet and IP-Adapter.
4. Perform quantitative and qualitative comparison between the improved pipeline built in #3 with the chosen baselines to determine the advantages of the proposed novelty.

1.3 Report Organisation

There are 5 chapters in this report, each chapter focusing on the following areas as below:

Chapter 1 (Introduction and Overview of the project): Introduction to the project's problem statement and intended outcomes.

Chapter 2 (Literature Review): Analysis of previous works done and shortlisting of state-of-the-art models for further analysis and benchmarking.

Chapter 3 (Analysis and Design of Proposed Approach): Introduction of the chosen approach and its' comparison of the proposed approach with chosen benchmarks.

Chapter 4 (Analysis and Discussion of Results): Analysis and discussion of results for all experiments performed.

Chapter 5 (Conclusion and Future Work): Conclusion and possible future improvements.

2 Literature Review

This section discusses previous work done for the task of Human Pose Transfer to identify state-of-the-art models used and possible modifications available for more sophisticated architectures.

2.1 Non-Diffusion Based Methods

In the early stages of development for the task of Human Pose Transfer, previous works involved Generative Adversarial Networks (GANs) [6] and Variational Autoencoders (VAEs) [7].

In GANs, two multilayer perceptron (i.e., generator *G* and discriminator *D*) are trained. The generator *G* is trained to learn a distribution p_g that provides a mapping function to represent a mapping of the input noise variables to p_z z a data space as $G(z;\theta_g)$. Meanwhile, the discriminator *D* is trained to maximise the probability of correctly identifying samples that came from the data rather than p_g. Together, the two perceptrons are simultaneously trained to maximise $log\ (1 - D(G(z))$. While the generation quality of GANs is high, the generations are constrained to the distribution of the inputs, and this may result in the generation not having much stochastic variation. This may be a limitation in the task of human image generation as generating human images may involve finer details in the subject's features (e.g., finer curls in hair) or style attributes. To overcome this, StyleGAN [8] proposes injecting noise after every convolution in the generator to introduce stochasticity. Also, a constant latent is passed into the generator instead with the input being passed through a series of FF networks before being injected into the generator via the AdaIN layers.

However, previous works involving GANs mainly focused on generating human faces and were still unable to generate full-body images of clothed humans. While GANs based methods may excel in reconstructing style attributes by blending colours, it is still lacking in their ability to move and deform objects which is pertinent for generating human images in different poses. Hence, later architectures begin to incorporate VAEs to learn about the underlying unseen variables of human images ($\boldsymbol{z}$) given the observed data ($\boldsymbol{x}$) for a choice of parameters ($\boldsymbol{\theta}$). In the context of Human Pose Transfer, VAEs aim to predict the appearance of humans in a human image involving an unobserved continuous random variable $\boldsymbol{z}$, that may depend on the style attributes and pose, through samples of observed data $\boldsymbol{x}$.

In VUNET [9], a conditional U-Net architecture with an additional VAE is proposed to condition the generation output on the latent representation of the source image and pose estimate produced by the VAE. This enables the U-Net generator to learn about the spatial properties of the image while the VAE learns about the style and appearance of the image. Together, these allow for the model to generate higher-quality images of humans in different poses with more intricate details. In HumanGAN [10], a VAE is used to extract the latent representation of the different human body parts (i.e., face, top, etc). These latent representations then have noise added and are warped with respect to a pose estimate before finally being passed to the generator.

2.2 Diffusion Based Methods

2.2.1 Diffusion Models

Despite remarkable advancement in generation qualities by GAN-based and VAE-based methods, GAN-based methods are trained to approximate the distribution of the input data thereby making it to mode collapse and training instability. Meanwhile, VAE-based methods are trained to learn latent variables through the observed input data using a set of parameters. However, the generation outputs of the VAE may be blurry as the training objective is minimised by taking an average over all possible outputs. Moreover, the chosen latent space dimensions may not be representative enough.

To improve on these, Denoising Diffusion Probabilistic Models (DDPM) [11] were introduced in 2020. DDPM are diffusion probabilistic models that is a parameterised Markov Chain trained using variational inference to produce samples matching the data after finite time which consists of the forward and reverse process.

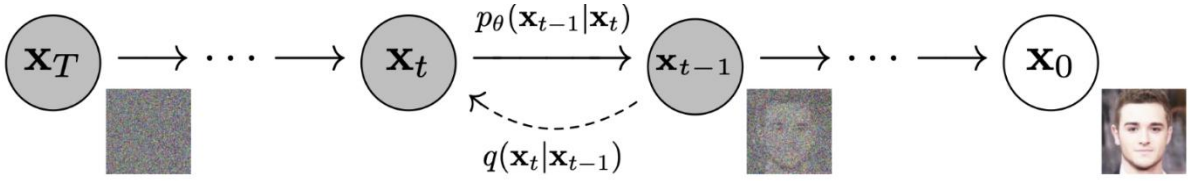


Fig 1: An overview of the forward, $q(x_t|x_{t-1}$, and reverse, $p_\theta(x_{t-1}|x_t$, process of the diffusion process where x_T is derived after adding noise to the original image, x_0, for T timesteps. Taken from [11].

In the forward (i.e., diffusion) process, starting from the original data at timestep t=0, the Markov chain gradually adds noise to the data until timestep t=T where the initial signal is destroyed, and the resulting data follows the standard distribution. Given the number of timesteps T and the variance schedule β, the approximate posterior $q(x_{1:T}|x_0$ is fixed to a Markov chain that gradually adds Gaussian noise. Given the observed data x_0, the forward process models the probability of observing $x_{1:T}$.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

Fig 2: Approximate posterior modelled in the forward process, where the variance schedule β determines the amount of noise added at each timestep. Taken from Equation (2) from [11].

In the reverse process, the joint distribution $p_\theta(x_{0:T})$ is defined as a Markov chain with learned Gaussian transitions starting at $p\ x_T\ := N(x_T; 0, I)$. Given the latent at timestep T, the reverse process models the probability that x_T was generated x_0 after T timesteps of adding noise (i.e., the probability of jointly observing $x_0, x_1, \ldots, x_T$).

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{1}$$

Fig 3: Joint distribution modelled in the reverse process, where $\mu_\theta(x_t|t)$ and $\Sigma_\theta(x_t|t)$ refers to the parameters of the Gaussian distribution at timestep t estimated by the diffusion models. Taken from Equation (1) from [11].

To efficiently generate images from this diffusion process, a model is trained to learn the amount of noise added between different timesteps (i.e., Gaussian transitions). This eventually allows for the prediction of the original data, x_0 given x_T which is sampled from the standard normal distribution. Hence, the training objective is to minimise the negative log-likelihood of $-\log(p_\theta\ x_0\ )$ with θ being the parameters of the diffusion model, which maximises the likelihood that the model can predict x_0 as accurately as possible from the diffusion process.

However, since the marginal likelihood $p_\theta(x_0)$ is intractable, training is performed by optimising the variational bound instead which encourages the final noisy sample to follow the standard Gaussian distribution and to minimise the difference in the predicted noise in the reverse process and the actual noise added in the forward process.

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] =: L \quad (3)$$

Fig 4: Training objective for diffusion models, where optimisation is performed by optimising the variational bound on the negative log likelihood. Taken from Equation (3) from [11].

2.2.2 Latent Diffusion Models

While diffusion probabilistic models achieved state-of-the-art results in sample quality, these models worked in the pixel space which brought about low inference speed and very high training costs. To achieve samples of higher quality, more denoising steps had to be performed which resulted in longer inference time. At the same time, training on high-resolution image data resulted in expensive calculations of gradients.

To overcome these shortcomings, the Latent Diffusion Model (LDMs) was proposed by [2] where the diffusion process works on the compressed latent space of the data of lower dimensionality. To further condition the generation output, conditioning using text prompts was also introduced in [2] using the cross-attention mechanism in the U-Net backbone thereby allowing for multi-modality conditioned image generation.

2.2.3 Developments in Diffusion Models for Human Pose Transfer

While Stable Diffusion has managed to achieve impressive in multi-modality (text and image) image generation, there still is much room for improvement in image generation within the context of Human Pose Transfer.

Specific to the generation of images in the context of Human Pose Transfer, information must be provided about the styles (i.e., fashion attributes) and the pose. This presents challenges as **(i)** the style and pose information must be integrated without entanglement to prevent potential conflicts between the conditioning provided and **(ii)** alternate modalities like image are more effective in describing the fashion attributes and the style attributes thereby requiring an effective way to condition the generation using multi-modality prompts.

2.2.3.1 Integrating Style and Pose information without entanglement

To address this issue, earlier studies performed in HumanDiffusion [12] and UPGPT [13] proposed to use cross attention in the attention layers within the blocks of the denoising U-Net to effectively incorporate conditionings of different modalities. These conditionings include the encodings of the context text describing the fashion attributes, the encodings of the fashion attributes from the style image, and the encodings of the pose.

Since the diffusion process operates in the pixel space, sophisticated alignment mechanisms may be needed when integrating multi-modality prompts (i.e., text and image in this case). Hence, while these methods try to promote disentanglement between the different conditioning by encoding each separately, adding or concatenating them before being passed as the query and key vectors in the cross-attention mechanism may not be sufficient.

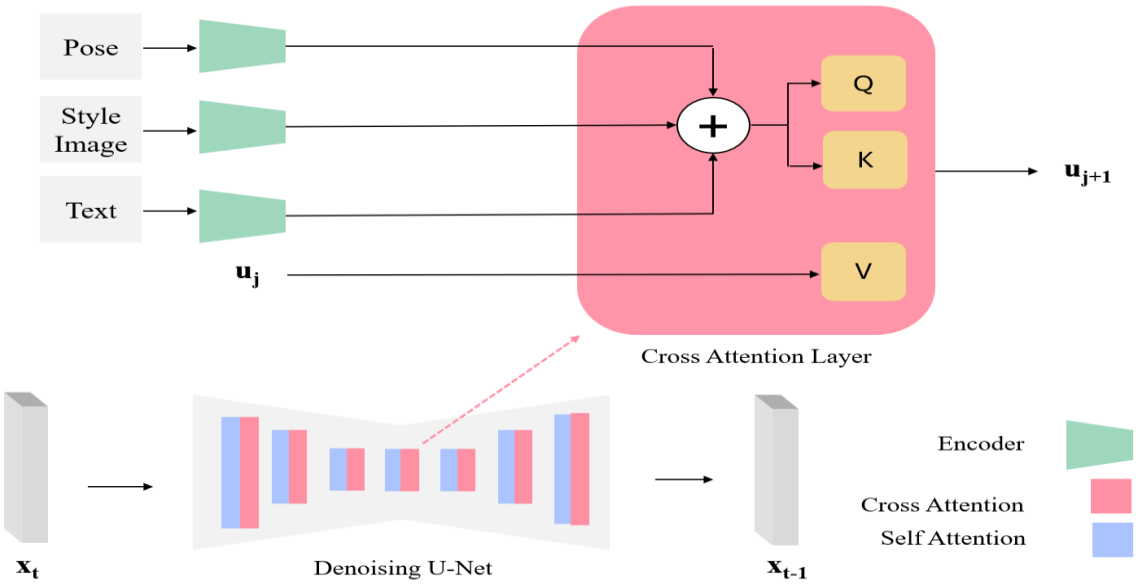


Fig 5. An overview of the method used to disentangle the different conditionings (e.g., pose, style image) using different encoders before passing it into the attention layers as seen in HumanDiffusion and UPGPT.

To reduce the entanglement between the different modalities, text conditioning was removed and model architectures focused on using image prompts to incorporate the fashion attributes from a style image and pose from a target image.

PIDM [14] was developed as a Denoising Diffusion Probabilistic Model with its outputs conditioned on a style image and pose from a target image. The pose was concatenated with the noisy target image while the style image was encoded by a texture encoder and passed into texture attention blocks in the denoising U-Net. Similarly, in Coarse-to-Fine Latent Diffusion (CFLD) [15], the style image was encoded to different resolutions and passed

through a set of learnable queries to get embeddings of the different fashion attributes. These embeddings were then used as the K and V vector of the cross-attention block while the pose was encoded using a series of convolutions and concatenated with the output of every encoder block of the U-Net. The removal of text conditioning could be seen as a sound choice as text conditioning may be less effective in describing stylistic attributes. Moreover, without a supplicated mechanism of performing multi-modality alignment, handling only one modality in the diffusion process may result in better performance. Nevertheless, removing text conditioning limits the expressiveness of the image generated. For example, text conditioning remains useful in describing specific niche styles (e.g., ukiyo-e style) or styles not necessarily observed in the training data.

2.2.3.2 Harmonising multi-modality conditional prompts

Given the image prompt's advantage in describing stylistic attributes, recent works in improving the conditional generation using both image and text prompts are highly relevant for the task of Human Pose Transfer. With previous works involving adapters gaining traction, this sparked interest in using lightweight adapters with the Stable Diffusion backbone to enable multi-modal conditional prompts to condition the generation at the expense of a modest increase in computational resources required.

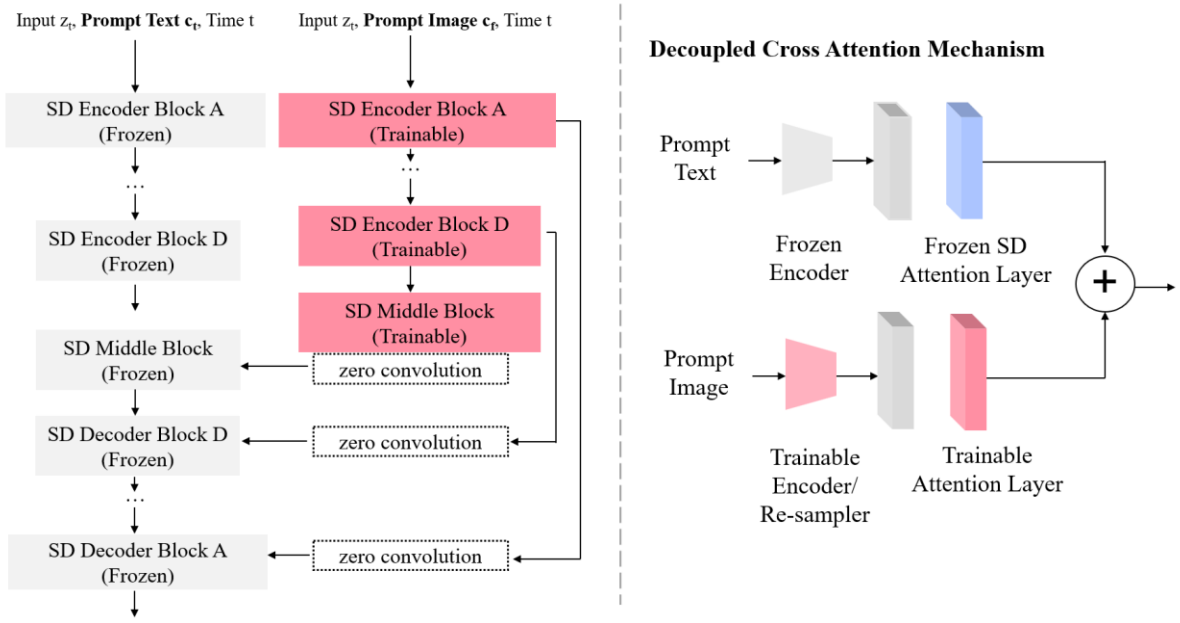


Fig 6. Model architectures proposed, ControlNet (left) and IP-Adapter (right), to incorporate image prompts in the text-to-image Stable Diffusion backbone.

In recent works, ControlNet introduced the concept of parallel branches to further condition the generation output of LDM with image prompts. Built on top of Stable Diffusion, an image prompt is passed through a trainable copy of the input blocks of the denoising U-Net and zero convolutions before being concatenated with the original branch's output. By passing each modality into a separate branch, each latent space receives only contextual conditioning from one modality at a time, thereby ensuring proper disentanglement between the modalities. Meanwhile, IP-Adapter, also built on top of Stable Diffusion, introduces the concept of decoupled cross attention where cross attention with the denoising U-Net's output at each layer is performed separately with the text and image prompt and is subsequently summed up. These approaches allow for different modalities (i.e., both text and image) to be used effectively to condition the generation while ensuring disentanglement between the modalities. More importantly, these novel adapter-based approaches are computationally efficient as well and improvements in conditional generation using multi-modal conditional prompts can be implemented without a significant increase in trainable parameters while maintaining the generation capabilities of the base Stable Diffusion model.

Following the advent of adapter-based approaches for multi-modal conditioning, Visconet [16] proposed a novel model architecture to further harmonise text and image conditioning using a single ControlNet. To ensure disentanglement between the different modalities, the style and pose conditioning were passed into the ControlNet while the text conditioning was passed into the original Stable Diffusion's branch. Most importantly, by processing each modality in a different branch, the conditioning strength of each modality can then be varied. This would allow for the image generated to contain a controllable amount of information from each modality by modifying the "control strength" parameter.

Fig 7. An illustration showing the effects of varying control strength. The outputs of Visconet are compared with ControlNet and IP-Adapter and suggests that Visconet can escape mode collapse faster and generate images with a harmonious image style while maintaining reasonable performance. These results are taken from [16].

Meanwhile, processing each modality separately and varying its influences can also help in overcoming the domain gap present within each modality and may prevent mode collapse. For example, conflicting words in complicated text prompts like “Ukiyo-e” and “Khaki” may result in mode collapse. With the architecture proposed by Visconet, each conflicting word may be represented using a different modality with its influence controlled by the “control strength” parameter.

Hence, this makes Visconet a promising option to perform Human Pose Transfer that ensures that complex stylistic attributes and pose can be incorporated into the generated image using both image and text prompts. This presents opportunities to generate complicated stylistic attributes that may be difficult to express with a single modality or even varying the background using descriptive text prompts without affecting the stylistic attributes of the subject without the need for much larger datasets covering all cases.

3 Analysis and Design of Proposed Approach

3.1 Analysis of Literature Review

Based on the literature survey conducted, earlier diffusion-based methods for Human Pose Transfer focused on developing architectures that were able to integrate the various information (i.e., pose, style, and text) without entanglement. In some related works such as PIDM and CFLD, text conditioning was removed and only image conditioning was utilised to condition the generation output based on pose and style.

However, the removal of text conditioning may limit the model's generative ability to perform style transfer or even generate new styles that were not present in the samples it was trained on. For example, text conditioning may be used to reinforce intricate styles present in the image conditioning, describe styles that were not seen in the training samples, or even generate illustrative backgrounds that match the fashion attributes of the subject.

Fortunately, recent developments in lightweight adapters allowed for reasonable performance in multi-modality image generation. This provides promising opportunities to include more modalities in describing the subject's fashion attributes or pose. Together with appropriate methods to prevent entanglement between information, this allows for information of different modalities to be used in conjunction without entanglement to provide improvements in generation outputs.

In Visconet, a novel architecture is proposed to harmonise both text and image prompts using a single ControlNet whereby the conditioning styles and pose are applied through the ControlNet branch while the conditioning text is applied through the original Stable Diffusion backbone. Using image prompts alone, Visconet has demonstrated reasonable performance in Human Pose Transfer. Even then, there remains an advantage that Visconet has over previously benchmarked works. With text prompts as an additional modality, this presents opportunities to generate more sophisticated and diverse fashion attributes or backgrounds.

Moving forward, the main objective of future works in this report would be to further improve on the results of Visconet while using it as a baseline.

3.2 Preliminaries

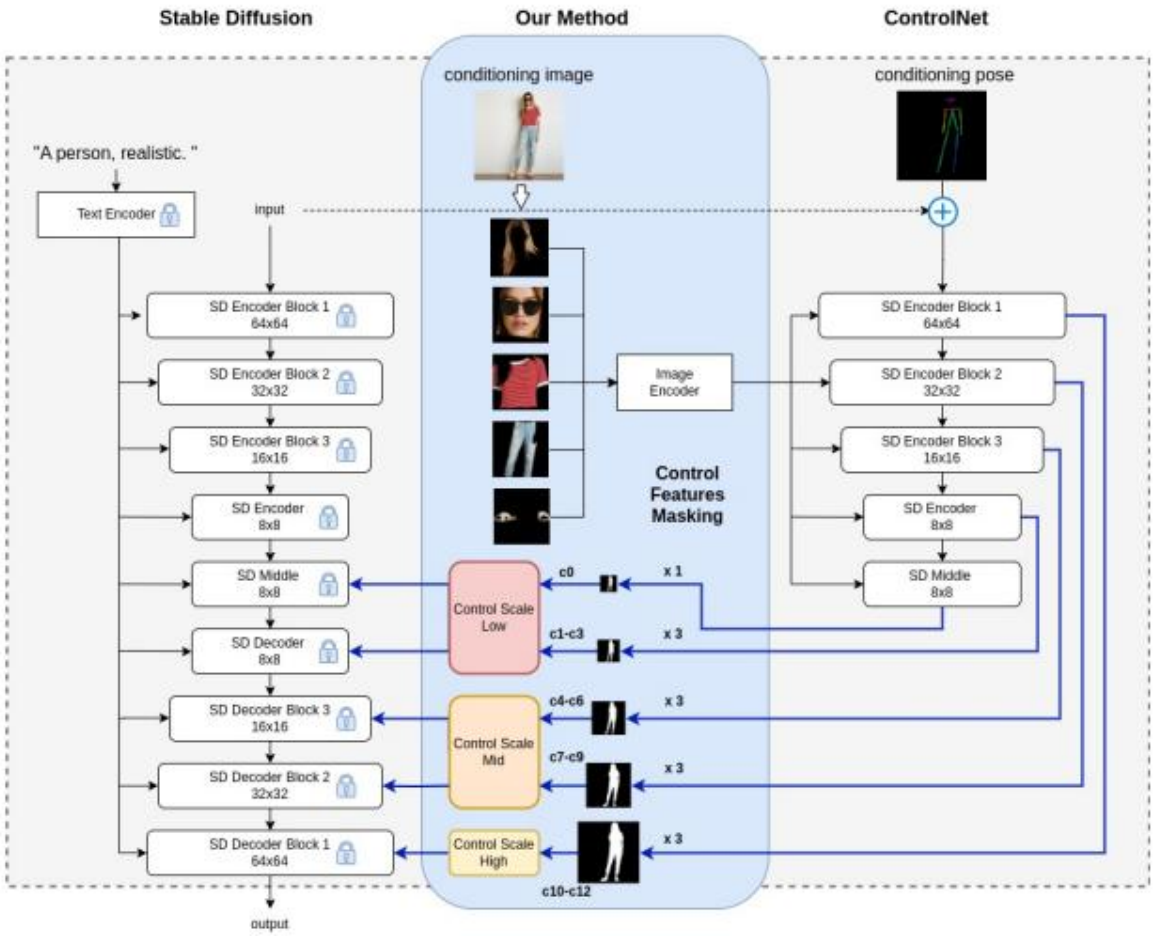


Fig 8. Architectural diagram of Visconet which utilises the Stable Diffusion Latent Diffusion Model conditioned on text as the backbone and an additional ControlNet branch. Time embedding, zero convolution and some blocks from the ControlNet diagram were removed for simplicity.

In the original Stable Diffusion, a U-Net [17] is used as the denoising network where input noise is progressively refined into latent variables which can be reconstructed into realistic synthetic images by learning the intricate distribution of images. The text conditionings are decomposed into smaller subunits, tokenised, and encoded with a CLIP [18] text transformer [19] to generate embeddings.

These embeddings are then injected into the cross-attention layers of the U-Net, serving as the sole conditioning in the generation progress. The loss function of the LDM is:

$$\mathcal{L}_{MSE} := \mathrm{E}_{z,c,t,\epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right]$$

where c is the text conditioning token, t is the diffusion time step, and z is the latent variable.

3.3 Visconet's main contributions

3.3.1 Replacement of Text with Visual Prompt

To allow for the harmonisation of text and image prompts, a ControlNet branch was utilised to handle all human visual appearances (both pose and style).

Previously, the conditioning used in the cross-attention layers of the ControlNet branch was text conditioning. However, Visconet replaced the text conditioning with visual conditioning concerning the fashion attributes of the subject. This prevents mode collapse in the ControlNet when contradicting conditionings are given in the text and image prompt in the ControlNet branch.

To generate the embeddings for the image conditioning, fashion attributes were first segmented from the source image and encoded using a CLIP vision transformer (*openai/clip-vit-large-patch14*) [20]. The embeddings were then pooled using a linear layer into length N (N=8 in Visconet) and are finally concatenated. The linear layer consists of only 2K parameters. However, given the intricacies of the fashion attributes (i.e., shape and texture), using a linear layer alone may not be enough.

3.3.2 Control Feature Masking

A binary human silhouette mask is applied to the control signals from the ControlNet branch before they are injected into the Stable Diffusion backbone. Applying this binary mask ensures that the style conditionings concerning the human visual appearance do not leak into the background. At the same time, it ensures that Stable Diffusion's generative capabilities are not severely restrained which may result in overfitting. This is because the dataset to be used, DeepFashion, consists of plain studio backgrounds which is vastly different from what Stable Diffusion was pretrained on. With masking applied, the loss function is now:

$$\mathcal{L}_{MSE} := \mathrm{E}_{z,c,t,\epsilon \sim \mathcal{N}(0,1)} \left[\mathcal{M} \odot \|\epsilon - \epsilon_\beta(z_t, t, c, v)\|_2^2 \right]$$

where ϵ_β is the model, $\odot$ is element-wise multiplication and $\mathcal{M}$ is the binary mask resized to the resolution (H,W) of the LDM output. Although only image conditionings are used in the model, text conditionings are used by the LDM in training hence, included in the equation.

3.4 Proposed further improvements

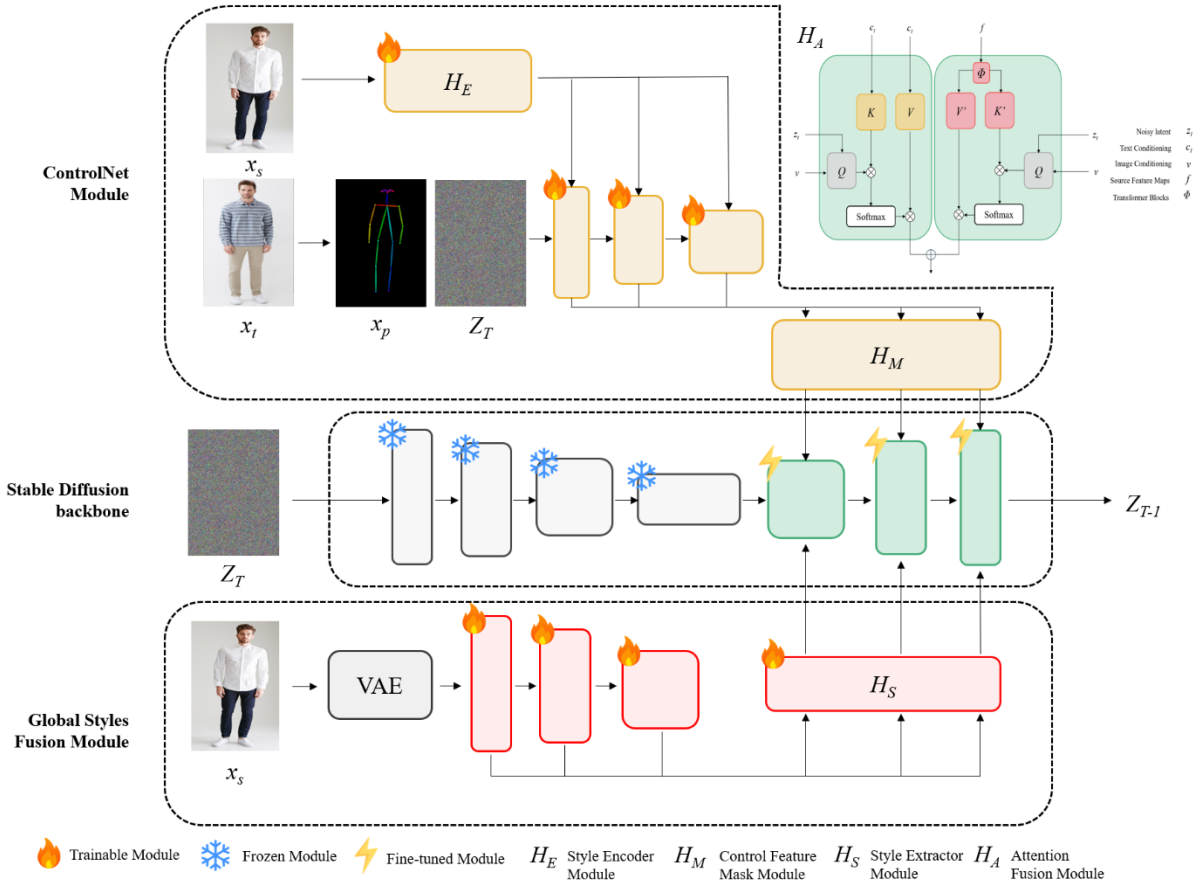


Fig 9. Overall architectural diagram of the proposed improved model. The model consists of two modules, ControlNet Module and Global Styles Fusion Module, built upon the Stable Diffusion backbone. Although not explicitly represented, the Global Styles Fusion Module encapsulates both the Style Extractor Module H_S and the Attention Fusion Module H_A.

The proposed architecture comprises the ControlNet module and the Global Styles Fusion Module built upon the Stable Diffusion backbone. Consisting of the Style Encoder Module H_E and the Control Feature Mask Module H_M, the ControlNet module aims to generate meaningful embeddings of the fashion attributes present in the style image and ensure that these embeddings are only applied to a specific region of the generated image (i.e., the foreground containing the subject) by applying control feature masking.

Meanwhile, the Global Styles Fusion Module consists of the Style Extractor Module H_B and Attention Fusion Module H_A. While the ControlNet module generates more local embeddings of the fashion attributes by performing piecewise or region-wise segmentation, the Global Styles Fusion Module aims to supplement these local embeddings by passing in the refined latent of masked human in the style image in its entirety as bias via the attention layers of the decoder blocks in the denoising U-Net present in the Stable Diffusion backbone.

Together, these modules aim to provide the model with a holistic representation of the styles present in the style image while retaining the backbones' generative capabilities with text prompts. Further elaboration on each module is described in the subsequent sections.

3.4.1 ControlNet Module

3.4.1.1 Style Encoder Module H_E

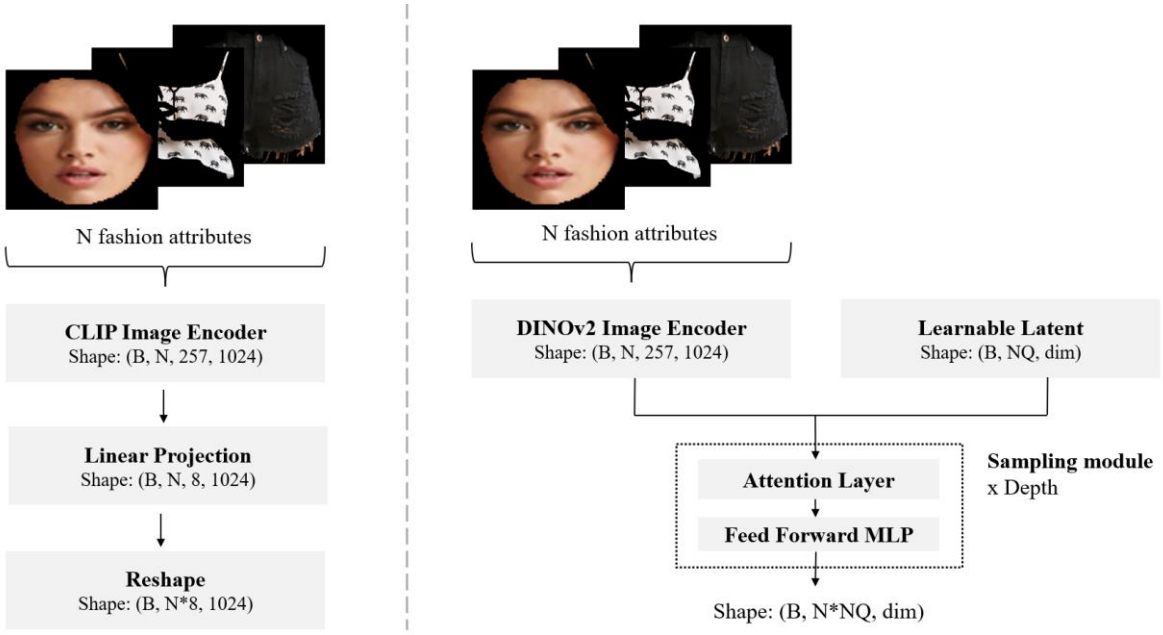


Fig 10. Model architecture of the style encoder used in Visconet (left) and the model architecture of the proposed style encoder in the improved Visconet (right). B refers to the batch size, N refers to the number of fashion attributes, NQ refers to the number of learnable queries and dim refers to the embedding dimensions.

The Style Encoder Module was implemented to provide an improved method to encode the fashion attributes present in the style image. While the intuition is largely like the style encoder implemented in the original Visconet, the improved Style Encoder Module proposes a different choice of the pre-trained image encoder and method used to generate a learnable set of queries to represent each fashion attribute.

In the improved image encoder for the fashion attributes, the pre-trained CLIP image encoder would be replaced with the pre-trained DINOv2 image encoder. In recent works, DINOv2 has been found to perform better at fine-grained detail recognition. This would be useful in identifying subtle details within fashion attributes thereby allowing for better preservation of styles. Moreover, unlike CLIP, DINOv2 was not specifically designed for text-image alignment tasks, which can be advantageous in purely visual tasks.

Taking inspiration from the Resampler of IP-Adapter Plus, a sequence of trainable parameters, with sequence length *NQ* and dimension *dim* is introduced. Each token in the sequence also referred to as queries, seeks to learn about different aspects of the fashion attributes (i.e., hair, face, etc). The interactions between the embeddings of the fashion attributes produced from the pre-trained image encoders and the set of learnable queries are then learned and exploited through a series of sampling modules, determined by the *depth* parameter to finally produce the final embeddings of the fashion attributes. By varying sequence length *NQ* and *depth* as parameters, the complexity of the model can be fine-tuned to ensure that the intricacies of the fashion attributes are well-captured in the embeddings while ensuring that redundancies is minimised.

3.4.1.2 Control Feature Mask Module H_M

The Control Feature Mask Module was implemented to provide an improved method to generate the mask based on the target image re-generation. Given a source-target image pair, the intended outcome is to generate an image of the subject in the target pose with the fashion attributes present in the source image. To ensure that no conditioning leaks into the background which may result in unwanted artifacts, it is important that the mask is sensitive to the pose and the fashion attributes the subject is wearing.

Currently, Visconet's implementation lacks a reliable method to generate the control feature mask. Currently, the control feature mask used is the human mask of the target image. In certain cases, different fashion attributes present in the source and target image may result in a mismatch in the control feature mask.

Fig 11. An illustration of unwanted artifacts being generated (red box) due to a mismatch in the control feature mask. The first row shows a positive example where there is no mismatch in the mask while the second row shows a negative example when there is a mismatch in the mask.

Ultimately, this results in unwanted artifacts being generated in areas, as seen in Fig 11. To overcome the previously mentioned weakness in Visconet's control feature masking, this work proposes an improved method based on previous works done in Virtual Try-On models.

In Virtual Try-On models, the aim is to simulate how a person would look wearing different outfits, accessories, and apparel in a given pose. In previous works for the task of Virtual Try-On models, the given inputs are an image of the warped cloth mask, the preserve image, and the pose conditioning generated from OpenPose [21] or DensePose [22] and the in-shop cloth.

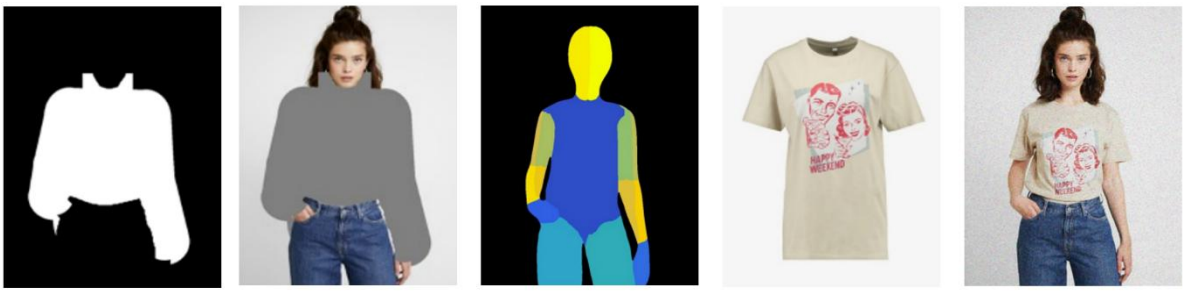


Fig 12. An illustration of the inputs used in Virtual Try-On models. These images are taken from IDM-VTON [23].

While the preserve images give hints on the fashion attributes to be preserved, the warped cloth mask gives hints on the shape of the area the model needs to re-generate using the in-shop cloth. In most cases, the shape of the warped cloth mask does not exactly match the shape of the in-shop cloth. This then requires the model to leverage its generative capabilities to generate accordingly in the area outlined by the warped cloth mask.

Fig 13. A step-by-step illustration of the generation of a coarser segmentation mask from the original mask derived from the target image.

Taking inspiration from Virtual Try-On models, the aim is then to perform a coarser segmentation of the subject in the target image. This allows the model to learn which part of the mask to generate as a fashion attribute or background when given a larger mask. At the same time, it allows for the smoothing of conditions applied along the borders of the mask, thereby resulting in a smoother transition between foreground to background.

After deriving the original mask from the target mask using the pre-trained fashion segment model *mattmdjaga/segformer_b2_clothes* on Hugging Face [24], the mask is then dilated using a square kernel of size 5 for 8 iterations. Then, the mask is blurred using a square kernel size of 5. Finally, the edge of the mask is smoothed using a transformation based on the distance from the border of the mask whereby the pixel value is determined as:

$$h_{mask} = 1 - \ e^{-k \bullet g\ d, d_{max}} \quad where\ g\ d, d_{max}\ = clip(\frac{d}{d_{max}}, 0,1)$$

3.4.2 Global Styles Fusion Module

While the introduction of the fashion attributes' encoding in the ControlNet branch aids in the preservation of style from the source image, the generated images may not contain the styles in their entirety (i.e., patterns and textures). This is because (i) the segmented fashion attributes obtained from the labelled segmentation mask or pre-trained fashion segmentation model may be at an angle depending on the subject's pose and (ii) some fashion attributes may be occluded by other more prominent fashion attributes when encoded into the same latent space by the encoder. In Fig 14, the first row illustrates a positive example where each

segmented attribute is well-represented and not warped. The second row and third row illustrate negative examples where some segmented attributes are occluded or when the pose results in warped segmented attributes that become occluded respectively. To resolve these issues, the injection of the latent of the source image at each dimension in the decoder is proposed.

Fig 14. An illustration of the fashion attributes (i.e., face, top, bottom, footwear) segmented from the source images. The first row illustrates a positive example while the last two rows illustrate negative examples.

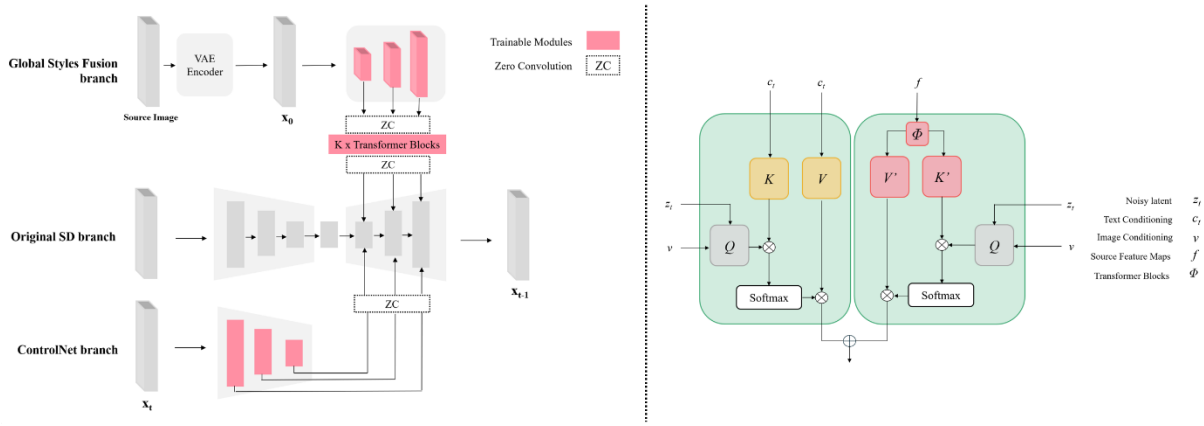


Fig 15. An illustration of how the latent of the source image is injected into the decoder via the Style Extractor Module H_S (left) and how it is used to condition the generated outputs in the decoder's attention blocks via the Attention Fusion Module H_A (right).

3.4.2.1 Style Extractor Module H_S

The latent of the source image is first passed into the pre-trained VAE encoder. Then, feature maps of various dimensions *[f_1, f_2, f_3]* are received using a feature map encoder *f*. Given latent of shape (*1,C,H,W*), where *C* is the model channels and (*H,W*) is the original dimensions of the latent derived from the pre-trained VAE encoder. The goal is to generate a more compact feature map of shapes (*1, mC, H//m, W//m*) where $m = 2^i$ and i is the depth of the decoder block. Then, the derived feature maps are passed through *K* transformer blocks, with zero convolution added in the beginning and the end, to dynamically determine the most important features of the latent to be used as conditioning in the cross-attention layers.

The zero convolution is a 1x1 convolution layer with both weight and bias initialised as zeros. This ensures that the learnt generative capability of the pretrained decoder is not negatively affected during the initial training phase.

3.4.2.2 Attention Fusion Module H_A

At every depth of the denoising U-Net decoder, each with different dimensions, the refined latent from the style images are then passed through a trainable copy of the K and V matrices of the cross-attention layers of the Attention blocks present in the denoising U-Net decoder, referred to as the keys *K'* and values *V'* matrices. In the original IP-Adapter [5], the style image in its entirety would be passed through an image encoder and embeddings are generated for the image in a global fashion. While this may allow for the overall global styles of the image to be preserved, this may not be ideal for the task of Human Pose Transfer where more fine-grained details regarding the subject's style need to be captured. Hence, by passing in refined latent from the style images at varying dimensions in the corresponding decoder's depth, it is believed that this would allow for the extraction and generation of features of different granularities from high to low-level features.

$$Z^{new} = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V + Softmax(\left(\frac{Q(K')^T}{\sqrt{d}}\right)V'$$

$$where\ Q = ZW_q,\ K = c_tW_k, V = \ c_tW_v,\ K' = \Phi(f)W'_k, V' = \Phi(f)W'_v$$

Moreover, to maximise computational efficiency, a parameter-efficient fine-tuning technique used for large models such as Stable Diffusion, LoRA: Low-Rank Adaption of Large Language Models [25] is used to represent the trainable copy of the keys *K* and values V

matrices of the cross-attention layers. Using two smaller trainable matrices with ranks that are much smaller than the original ranks of the matrices present in the attention layers, this allows for the increase in performance with the minimal increase in the number of trainable parameters.

$$W'_k = W'_{B,k} W'_{A,k}\,,\; W'_v = W'_{B,v} W'_{A,v}$$

where $W'_{B,k}$ is the reconstruction matrix of W'_k with shape $[D, R]$ and $W'_{A,k}$ is the decomposition matrix of W'_k with shape $[R, D_f]$. *D* refers to the inner dimension at which the attention mechanism operates at, *R* refers to the rank of the LoRA matrices and D_f refers to the embedding dimension derived from the Style Extractor Module H_S derived at each dimension. Together, applying the decomposition and reconstruction matrix sequentially reconstructs the original linear mapping in W'_k but with a reduced number of trainable parameters.

3.4.3 Fine-tuning of the Decoder for smoother backgrounds

Finally, the cross-attention layers in the decoder's attention blocks would be fine-tuned using the same parameter-efficient fine-tuning technique used for large models such as Stable Diffusion, LoRA: Low-Rank Adaptation of Large Language Models. While the Stable Diffusion backbone's generative capabilities allow it to generate a diverse range of illustrative backgrounds, the model has trouble generating plain simple backgrounds observed in the DeepFashion MultiModal dataset. This often results in unnatural distinct patches of colours in the background. To resolve this, LoRA fine-tuning is adopted by introducing additional fine-tuning parameters for the Q and V vectors of the cross-attention layers. Setting a constant text prompt of "a person. plain studio background" and performing fine-tuning, this allows the model to learn how to better align the noisy latent in the Stable Diffusion backbone with the text conditioning of a simple studio background. A scaling alpha parameter α ranging from 0 to 1 can be applied to control the conditionings applied from the LoRA matrices. Hence, the final output from the cross-attention layer in the attention block can be calculated as

$$Z^{new} = Softmax\left(\frac{Q'K^T}{\sqrt{d}}\right)V',\; Q' = Q + \alpha_Q W'_{B,Q} W'_{A,Q}, V' = V + \alpha_V W'_{B,V} W'_{A,V}$$

where α_Q is the alpha scaling parameter for *Q* matrix, $W'_{B,Q}$ is the reconstruction matrix of W'_Q with shape $[D, R]$ and $W'_{A,Q}$ is the decomposition matrix of W'_Q with shape $[R, D_{z_t}]$ and likewise for the *V* matrix. *D* refers to the inner dimension at which the attention mechanism operates at, *R* refers to the rank of the LoRA matrices and D_{z_t} refers to the dimensions of the noisy latent in the denoising U-Net backbone.

3.5 Other baselines to be considered

To qualitatively analyse the results, a few benchmark models, categorised under the following categories, would be used. These benchmarks are diffusion based state-of-the-art models that do not incorporate the use of adapters (PIDM and CFLD), a naïve implementation of Stable Diffusion with ControlNet and IP-Adapter, and finally, Visconet.

In the naïve implementation of Stable Diffusion with ControlNet and IP-Adapter, a source image containing the fashion styles to be incorporated is passed through the IP-Adapter's Image Projection Model where the derived style embeddings are subsequently used in the decoupled cross attention mechanism. Meanwhile, an OpenPose pose map is passed through the ControlNet branch to condition the generation based on the provided pose. A simple text prompt of "a person." would be given to avoid potential conflicts with the image conditionings.

For the Stable Diffusion backbone, the pretrained weights were loaded from *benjamin-paine/stable-diffusion-v1-5* [26]. For the ControlNet, the pretrained weights were loaded from *lllyasviel/control_v11p_sd15_openpose* [27]. For the IP-Adapter, the pretrained weights for both the IP-Adapter and IP-Adapter-Plus versions were loaded from *h94/IP-Adapter* [28].

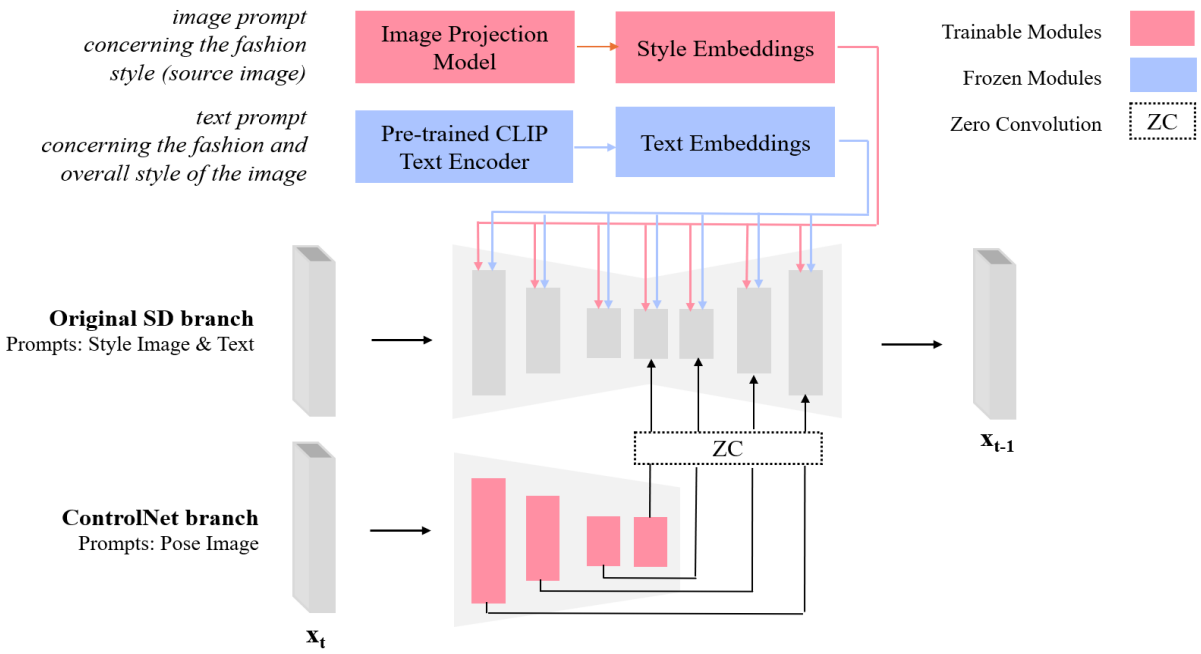


Fig 16. An overview of the model architecture combining both the ControlNet and IP-Adapter adapters with the Stable Diffusion (SD) backbone to perform the task of Human Pose Transfer.

3.6 Evaluation metrics

To evaluate the performance of each ablation, the following widely used quantitative evaluation metrics in human image generation are considered.

3.6.1 Structural Similarity (SSIM)

SSIM [29] is a non-network-based metric that is based on traditional image processing techniques and serves as a measure of similarity between the generated and reference images by considering factors such as luminance, contrast, and structure. Given generated image x and reference image y, μ and σ represents the mean and standard deviation of the respective images. Meanwhile, σ_{xy} represents the correlation coefficient between x and y. α, β and γ are parameters typically set to 1 and C_1, C_2 and C_3 are predefined constants. Together, the SSIM score is calculated by taking the product of the similarities of luminance $I(x,y)$, contrast $C(x,y)$, and structure $S(x,y)$ between the images. The SSIM value falls within the range of 0 to 1 with a larger value indicating a greater degree of similarity.

$$SSIM\ x,y\ = I\ x,y\ ^{\alpha}C(x,y)^{\beta}S(x,y)^{\gamma}$$

$$where\ I\ x,y\ = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, C\ x,y\ = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, S\ x,y\ = \frac{\sigma_{xy} + C_3}{\sigma_x + \sigma_y + C_3}$$

3.6.2 Fréchet Inception Distance (FID)

The Fréchet Inception Distance [30] is a network-based metric that serves as a metric for assessing the quality of images synthesised by generative models. By assuming that both the generated image x and reference image y conform to a Gaussian distribution, this metric quantifies the Fréchet distance between these distributions using a batch of real and generated images. Using a pre-trained inception network, features from both the real and generated images are extracted. In this work, the inceptionv3 feature layer chosen is 64. Then, the feature from each image is postulated to follow a multidimensional Gaussian distribution where $N(\mu_x, \Sigma_x)$ and $N(\mu_y, \Sigma_y)$ is the distribution of the generated and real image respectively. A lower FID score indicates that the distribution of the extracted features from both images is more similar.

$$FID\ x, y\ = \|\mu_x - \mu_y\|_2^2 + tr(\Sigma_x + \Sigma_y - 2\sqrt{\Sigma_x}\Sigma_y$$

3.6.3 Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [31] is another popular network-based metric that aims to quantify the perceptual distance between a given pair of generated image x and reference image y. Feature representation extracted from corresponding patches of each image is first extracted using a pre-trained CNN model. Then, LPIPS is calculated as the average of the similarity between the extracted features from all the image patches. Possible similarity metrics are cosine similarity or Euclidean distance. In this work, the VGG backbone and L2 similarity metric are chosen.

$$LPIPS\ x, y\ = \frac{1}{N}\sum_{i=1}^{N} d(F_i(x), F_i(y))$$

where N represents the number of image patches, $F(\cdot)$ denotes the extracted feature representation of the image patch and $d(\cdot)$ denotes the distance metric used.

4 Analysis and Discussion of Results

This section covers the datasets used, the training and inference configurations, and finally the results of each model iteration in comparison to the chosen benchmarks.

4.1 Datasets Used

DeepFashion [32] is a popular and de-facto dataset used for Human Image Generation in recent literature with over 800,000 diverse fashion images coupled with rich annotations. In particular, the dataset used in the following experiment is the DeepFashion MultiModal dataset [33] which is a subset of DeepFashion's In-shop Clothes Retrieval Benchmark.

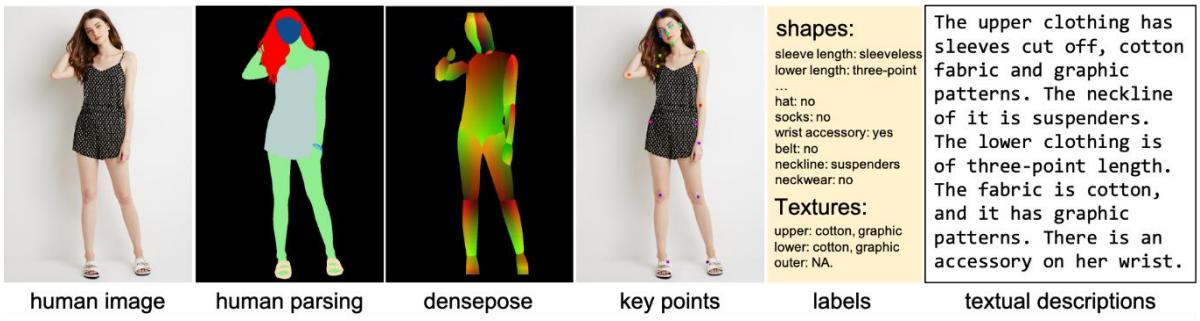


Fig 17. Overview of the DeepFashion MultiModal dataset.

More precisely, the DeepFashion MultiModal dataset contains 44,096 fashion images of a subject wearing a variety of fashion apparel on a plain, white background. Out of these images, 12,701 full-body human images were further subsetted to be used for training. This ensured the accurate derivation of the skeletal map from the human image and accurate segmentation of fashion attributes from the "human parsing" annotation.

4.2 Training Setup

To train the model, train-test-split was first performed on the 12,701 full-body human images present in the DeepFashion MultiModal dataset where 11,000 images were used for train and validation while the remaining images were used for testing. Then, at each ablation, a subset of the train dataset was used. This was done to ensure that training could be completed in a reasonable time while allowing for the opportunity to study the model's performance when trained on different sizes of the dataset. Train-test split was then performed on each gender's images using a ratio of 90:10. Finally, a predetermined number of random pairs were randomly chosen to be included in the final train and test dataset.

Pose information is extracted using OpenPose to create body-and-hand skeletal pose images and pre-segmented fashion attributes are taken from the “human parsing” annotation of the DeepFashion MultiModal dataset. A neutral and simple text prompt of “a person.” is used for all images to prevent conflicts with the Stable Diffusion backbone. At the same time, it acts as an unconditional text embedding thereby allowing users to amplify visual effects using positive prompts, negative prompts, and guidance scales.

Like other adapter-based models, pre-trained Stable Diffusion weights are used. In this work, experiments were performed using the pre-trained weights of Stable Diffusion v2.1 [34]. The adapter branch was initialised by copying frozen weights from the Stable Diffusion. All weights in the Stable Diffusion, CLIP text encoder, and image encoder are frozen. In the ControlNet branch, only the weights of the Style Encoder Module H_E, consisting of the Resampler and the trainable copy of the encoder blocks in the ControlNet branch are updated during training. Meanwhile, only the weights of the Style Extractor Module H_S and the weights of the key K' and value V' matrix in the Attention Fusion Module H_A are updated during training. For the image encoder, the CLIP image encoder *clip-vit-large-patch14* or the DINOv2 image encoder *dinov2-base* [35] are used. The remaining configurations were retained from ControlNet.

During training, 5% of the samples were conditioned with null inputs to encourage the model to develop stronger unconditional guidance capabilities. The model was trained across four GeForce RTX 3090 GPUs for at least 50,000 training steps, each with a batch size of 1 with gradient accumulation every 4 batches, effectively resulting in a batch size of 16 which aligns with Visconet’s configuration. Due to time constraints, training was stopped early when no significant gain in performance on the validation dataset was observed. Multi-GPU parallel training is performed with Pytorch Lightning’s Distributed Data Parallel strategy [36].

4.3 Experiments

This section details the various experiments performed.

4.3.1 Baseline Benchmarks

Fig 18. Comparison of results obtained from each chosen baseline model architecture. The first column shows the source image, the second column shows the conditioning pose while the remaining columns shows generated outputs from each baseline.

Model	SSIM (↑)	FID (↓)	LPIPS (↓)
PIDM	**0.656**	0.390	0.184
CFLD	0.660	**0.195**	**0.183**
Naïve HF implementation	0.566	7.25	0.234
Baseline Visconet	0.401	9.79	0.306

Table 1. Evaluation metrics for the chosen baseline benchmarks. Naïve HF implementation refers to the naïve implementation of Stable Diffusion with ControlNet and IP-Adapter on Hugging Face.

Preliminary results were first derived from the chosen baseline, namely from the (i) previously established works that did not utilise the combination of ControlNet and IP-Adapter (PIDM and CFLD), (ii) naïve Hugging Face implementation of StableDiffusion with ControlNet and IP-Adapter (Fig 14), to be referred to as naïve Stable Diffusion pipeline from here on, and (iii) baseline Visconet. To ensure a fair comparison, a text prompt of "a person." was used for both (ii) and (iii). This neutral and simple text prompt prevents conflicts with the Stable Diffusion backbone and ensures that the generations were only conditioned on image prompts, which were available for all baselines.

The generative capabilities of Stable Diffusion's LDM can be observed by comparing the generated outputs of PIDM, CFLD, and the naïve implementation of Stable Diffusion with ControlNet and IP-Adapter from Hugging Face. PIDM was trained on the DeepFashion In-Shop Clothes Retrieval Benchmark [37] and Market-1501 [38] dataset while CFLD was trained on DeepFashion' s In-Shop Clothes Retrieval benchmark. These datasets contain full-body human images with plain and dull backgrounds. Meanwhile, the Stable Diffusion LDM and adapters were trained on a much larger and diverse dataset. Moreover, PIDM and CFLD did not incorporate the use of any pre-trained image adapters which may have created a larger domain gap. Hence, there exists a domain gap when the naïve Stable Diffusion pipeline is asked to generate human images with a simple background. This results in unwanted fashion artifacts that were not present in the source image or unwanted artifacts in the background.

To overcome this weakness, the baseline Visconet has implemented masking of the control features to ensure that the generative capabilities of the LDM do not leak into the background and are applied only to the fashion attributes. However, it seems that the styles of the fashion attributes are not preserved in the baseline Visconet. As seen in the last row for baseline Visconet, hallucination also occurs when there is a mismatch in the source image and the control feature mask, obtained from the human mask of the target image.

4.3.2 Experiment 1 – Implementation of the Style Encoder Module H_E

During the training phase, the same image is used in each source-target pair and the model learns the process of reconstructing the source image given the pose and style conditionings.

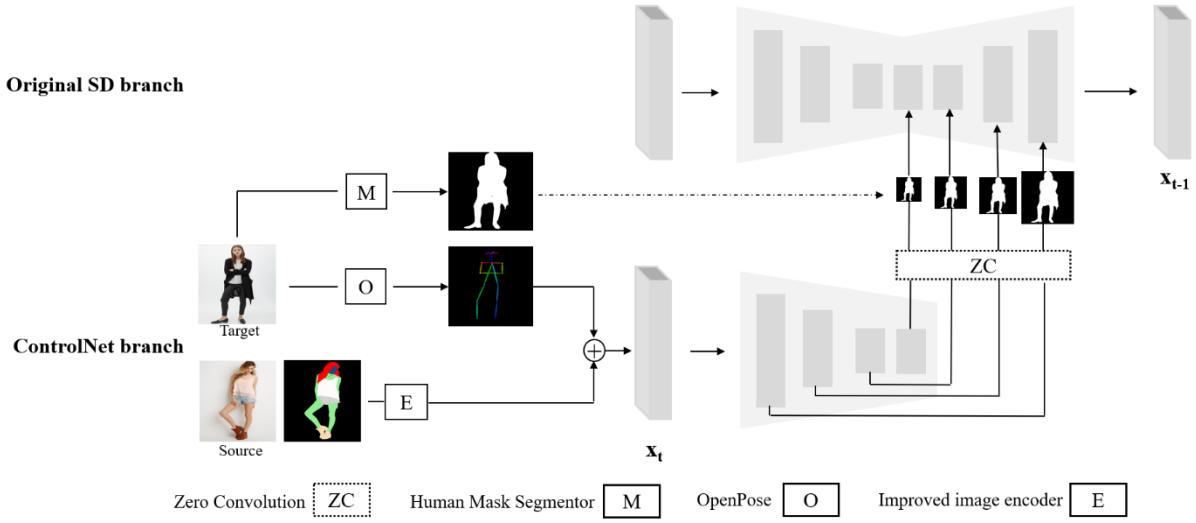


Fig 19. Detailed architectural diagram of the model used in the first experiment.

The target image is passed into a pre-trained segmentation model, specifically the DeepLabv3 segmentation model with a ResNet-101 backbone [39], available from the torchvision library to obtain the segmented human mask. This mask is then used as the control feature mask. Pose information is extracted from the target image using OpenPose annotations. Meanwhile, the fashion attributes present in the source image are first segmented using pre-segmented fashion images available from the DeepFashion Multimodal dataset. The segmented fashion attributes are then passed into the proposed improved image encoder to obtain the fashion style embeddings.

For brevity, each variant would be referred to as:

Baseline Visconet: Baseline Visconet

Model 1: Improved Visconet, CLIP Encoder with 16 queries

Model 2: Improved Visconet, CLIP Encoder with 24 queries

Model 3: Improved Visconet, DINO Encoder with 16 queries

The configuration YAML file for each model iteration can be found at [40], where Model 1 pertains to *visconet_v2_baseline_pair.yaml*, Model 2 pertains to *visconet_v5_pair.yaml*, Model 3 pertains to *visconet_v6_pair.yaml* and Model 4 pertains to *visconet_v7_pair.yaml*.

Number of fashion attributes	7
Target fashion attributes	Top, Outer, Skirt, Dress, Headwear, Rompers, Pants, Footwear, Hair, Face, Belt
Criterion	MSE loss
Optimiser	AdamW, with default parameters
Learning Rate	5×10^{-5}
Learning Rate Scheduler	ReduceLROnPlateau with patience = 3, threshold = 0.001, factor = 0.5, min_lr = 1×10^{-5}
Batch Size	16 (1 per GPU with gradient accumulation of 4)
Number of training steps	64,000
Training Dataset size	2,000 samples
Resumed Checkpoint	visconet_v1.pth

Table 2. Common configurations/hyperparameters used in Experiment 1. visconet_v1.pth refers to the initial pre-trained checkpoint of the baseline Visconet.

Fig 20. Comparison of results obtained from the baseline Visconet and the variants of the improved Visconet, which incorporates the use of a Resampler, inspired from IP-Adapter Plus, and the DINO v2 image encoder. Each pair of image shows the generated output from each model and the zoomed in area to be focused on.

Fig 21. Comparison of results obtained from the baseline Visconet and the variants of the improved Visconet, which incorporates the use of a Resampler, inspired from IP-Adapter Plus, and the DINO v2 image encoder. The first two rows show existing weaknesses in the current implementation method while the last row shows the best results attainable from the most 'desirable' source-target image pair.

Model	**SSIM (↑)**	**FID (↓)**	**LPIPS (↓)**
PIDM *(B)*	**0.656**	0.390	0.184
CFLD *(B)*	0.660	**0.195**	**0.183**
Naïve HF implementation *(B)*	0.566	7.25	0.234
Baseline Visconet *(B)*	0.401	9.79	0.306
Model 1	0.418	9.13	**0.297**
Model 2	0.411	9.49	0.300
Model 3	**0.424**	**8.426**	0.301

Table 3. Evaluation metrics for the models in Experiment 1 with the chosen baseline benchmarks. *(B)* denotes a baseline model.

Model	**ΔSSIM (↑)**	**ΔFID (↓)**	**ΔLPIPS (↓)**
	$SSIM_0 = 0.401$	$FID_0 = 9.79$	$LPIPS_0 = 0.306$
Model 1	0.016	-0.661	**-0.009**
Model 2	0.010	-0.295	-0.005
Model 3	**0.023**	**-1.36**	-0.005

Table 4. Improvement in metrics for each of Experiment 1's model variation in comparison to the Baseline Visconet (e.g., $SSIM_i - SSIM_0$). The metrics obtained for Baseline Visconet are labelled below each column header.

Fig 20 shows the results from the task of reconstruction, whereby the source and target image are the same. This was first done to indicate how well the improved architecture is at preserving the fashion attributes present in the source image, with all else held constant. Analysing these results, all improved variants of Visconet seem to be better able to preserve the styles of the fashion attributes. Colours, intricate patterns and designs are better represented in the generations. Comparing the generated outputs of the improved Visconet using the CLIP encoder with 16 and 24 queries respectively, it can be observed that the intricacies of fashion attributes can still be further captured. This motivates the need for each fashion attribute to be represented by a larger number of queries in the Resampler.

Next, Fig 21 shows the actual task to be solved, whereby the source and target images are different. From these results, two shortcomings of the current method can be identified. Firstly, the mismatch between the fashion attributes present in the source image and the shape of the target mask may result in unwanted artifacts being generated, as seen in row 1 of Fig 19. This also leads to an underrepresentation of fashion attributes present in the source image, as seen in row 2 of Fig 19 whereby the restrictions imposed by the target mask result in the loose-fitting flowy dress being represented as jeans. Moreover, obvious contrasts can be seen at the borders of the mask, between foreground and background.

Meanwhile, the second row of Fig 21 also demonstrates the need for a coarser-based segmentation of fashion attributes due to the possible occlusion of some fashion attributes. For example, instead of segmenting the “outer” and “dress” fashion attributes separately, both should be segmented as a single attribute for their embedding to be representative. Also, the model is unable to generate the correct fashion attributes. In this example, the model mistakes the combination of an “outer” and “pants” with a “shirt” and “pants”. While the model can generate cloth of the correct colour, it is unable to correctly generate the correct fashion attribute present in the source image. This is because the colour and texture of the clothing are well-represented using the fashion attributes encoder but only encoding segmented fashion attributes piecewise renders no global information to the model about the fashion attributes originally present in the source image.

Generally, model variations that are adapter-based methods with a Stable Diffusion backbone (from naïve HF implementation onwards) tend to exhibit poorer performance, especially in network-based approaches that compare the difference in feature distributions. Observing the generated images for each model variation, the unwanted artifacts and coloured backgrounds

are major causes of the poor performance observed. Hence, future experiments are targeted at overcoming these pertinent issues. Given that Model 3 (Improved Visconet, DINO Encoder with 16 queries) exhibits the best improvements in metrics out of all model variations in Experiment 1, it would be considered the leading model for Experiment 1 in future ablations. Future experiments would also be built upon it.

4.3.3 Experiment 2 - Implementation of the Control Feature Mask Module H_M

In the previous ablation, the segmented human mask of the segmented image was used as the control feature mask. However, this resulted in unwanted artifacts being generated in the background due to a mismatch in the shape of the control feature mask. At the same time, details about the background were added to the text prompt to ensure a smoother transition from foreground to background. To overcome these issues, the improved masking pipeline introduced in Fig 13. is employed in this experiment.

It was also observed that a coarser segmentation of the fashion attributes is required to prevent the possible occlusion of some fashion attributes.

The remaining hyperparameters remain the same as in Experiment 1, as detailed in Table 1. Meanwhile, Table 2 details the most pertinent hyperparameters used in the improved masking pipeline introduced in Fig 13.

Number of fashion attributes	4
Target fashion attributes	Face (hair, face, headwear), Top (top, outer, dress), Bottom (skirt, dress, pants, leggings), and Footwear
Dilate Kernel Size	5x5
Dilate Iterations	8
Blur Kernel Size	25
Maximum Decay Distance	30
Decay Scale	Exponential
Decay Scale Factor (Rate of decrease)	3
Number of training steps	64,000
Training Dataset size	2,000 samples
Resumed Checkpoint	visconet_v1.pth

Table 5. Hyperparameters used in the improved masking pipeline. visconet_v1.pth refers to the initial pre-trained checkpoint of the baseline Visconet.

For brevity, the improved Visconet model variations used in this ablation are:

Model 3: Improved Visconet, DINO Encoder with 16 queries (Ablation 1, Model 3)

Model 4: Improved Visconet, DINO Encoder with 16 queries, Improved Masking, Text Prompt of “a person.”

Model 5: Improved Visconet, DINO Encoder with 16 queries, Improved Masking, Text Prompt of "a person. plain studio background."

Fig 22. Comparison of results obtained from the original masking approach with the variants incorporating the improved masking pipeline.

Model	SSIM (↑)	FID (↓)	LPIPS (↓)
PIDM *(B)*	**0.656**	0.390	0.184
CFLD *(B)*	0.660	**0.195**	**0.183**
Naïve HF implementation *(B)*	0.566	7.25	0.234
Baseline Visconet *(B)*	0.401	9.79	0.306
Model 3 (*1)	0.424	8.426	0.301
Model 4	0.451	6.918	0.278
Model 5	**0.463**	**6.197**	**0.271**

Table 6. Evaluation metrics for the models in Experiment 2 with the chosen baseline benchmarks. *(B)* denotes a baseline model while (**i*) denotes the best performing model from the previous i^{th} experiment.

Model	ΔSSIM (↑)	ΔFID (↓)	ΔLPIPS (↓)
	$SSIM_0 = 0.424$	$FID_0 = 8.426$	$LPIPS_0 = 0.301$
Model 4	0.027	-1.508	-0.023
Model 5	**0.039**	**-2.259**	**-0.030**

Table 7. Improvement in metrics for each of Experiment 2's model variation in comparison to the best performing model variation from Experiment 1 (e.g., $SSIM_i - SSIM_0$).

In the previous experiments, the segmented human mask from the target human was used for the control feature masking. As a result of this precise masking, the model is conditioned to generate using the fashion attributes conditioning on every masked pixel and is not able to handle the generation of pixels that may pertain to the background. This results in unwanted artifacts being generated whenever there is a mismatch in the control feature mask applied (i.e., when the given mask occupies pixels concerning both the subject and the background).

With the improved control feature masking, the model is now able to discern which parts of the mask to generate as fashion attribute or backgrounds given a dilated and smooth mask. At the same time, this also smoothened out the differences between the foreground and background in some samples, further improvements can still be made as seen in the first two rows of Fig 20. Together, these improvements removed unwanted artifacts which has resulted in an improvement in the chosen metrics.

In hopes to further reduce the discrepancy between the background inside and outside of the mask, the text prompt was modified to "a person. plain studio background." However, empirical results suggest limited improvements. Instead, further ablations and training on a larger dataset may be required for the model to generate a more homogeneous plain background.

Finally, it can be observed that the model is still unable to correctly generate the correct fashion attribute present in the source image in certain samples, such as in the third row. As mentioned in Section 3.4.3, certain segmented fashion attributes are not well-represented due to the pose or more prominent fashion attributes. Hence, simply conditioning them into the same latent space during the encoder phase may not be sufficient in representing them. This suggests the need for the latent representation of the source image to be injected as bias in the decoder to provide the model with a more comprehensive understanding of the global composition of the image which will be explored in the next experiment.

While Model 5 from the Experiment 2 demonstrated good pose accuracy, it was still not able to replicate the styles present in the source image to its fullest. For example, from Fig 22, the black knee-long cardigan and white jumpsuit worn by the subject in column 4 were represented as a black top and black pants. Meanwhile, the inner grey T-shirt worn by the subject in column 5 was underrepresented in the final generated image. These negative examples suggest the unresolved challenge of representing occluded fashion attributes present in the style image. Moreover, the distinct patches of colour clearly differentiating the foreground and the background suggests unresolved difficulties in fusing the generation output between the regions conditioned and unconditioned by the ControlNet.

4.3.4 Experiment 3 - Implementation of the Global Styles Fusion Module

Experiment 3 intends to resolve the challenge of representing occluded fashion attributes present in the style image. In this experiment, the latent representation of the source image is first refined at various dimensions via the Style Extractor Module H_S before it is incorporated into the cross-attention layers in the decoder's attention blocks via the Attention Fusion Module H_A. Table 8 details the most pertinent parameters concerning H_S, H_A and the training set-up. The remaining training hyperparameters, choice of hyperparameter for H_E and H_M remains the same as Experiment 2.

H_S	transformer_block_dim	256
	transformer_depth	1
	get_featuremaps_method	"attn"
	proj_out_dim	77
	ip_mask_only	True

H_A	ip_rank	4
Train Related	Number of training steps	109,000
	Training Dataset size	11,000 samples
	Resumed Checkpoint	expt2-model5-gs155k.ckpt

Table 8. Hyperparameters used to train the Global Styles Fusion Module. expt2-model5-gs155k.ckpt refers to the pre-trained checkpoint from Model 5 of Experiment 2, up till a global step of 155,000.

To generate compact feature maps of the latent of the source image, two different methods of varying complexity, the feature map encoder *f* were implemented. A less complex convolution-based approach, referred to as *Conv*, was first experimented with. This method involved a series of Residual Blocks each consisting of a sequence of convolution, group normalisation, and sigmoid linear unit activation performed in sequence. Alternatively, a more complex transformer-based approach, referred to as *Attn* was experimented with too. This method involves the integration of Swin Transformer [41], a hierarchical vision transformer, to produce more meaningful feature maps using shifted windows. For brevity, only the experiment involving the transformer-based approach is presented here.

Finally, the *ip_mask_only* Boolean flag was varied as well. If set to true, only the masked human instead of the entire source image is passed as input to the pre-trained VAE encoder. In the DeepFashion Dataset, most images contain a plain studio background. While using the source image in its entirety would likely yield better performance for the task of pose transfer with the DeepFashion dataset, it may affect the model's ability to generalise to source images from different sources if the background is injected as bias too in the decoder. At the same time, this also prevents potential clashes with the text conditioning applied in the original Stable Diffusion branch.

Model	SSIM (↑)	FID (↓)	LPIPS (↓)
PIDM *(B)*	**0.656**	0.390	0.184
CFLD *(B)*	0.660	**0.195**	**0.183**
Naïve HF implementation *(B)*	0.566	7.25	0.234
Baseline Visconet *(B)*	0.401	9.79	0.306
Model 5 (*2)	0.463	6.197	0.271
Model 6 (*3)	**0.538**	**1.557**	**0.242**

Table 9. Evaluation metrics for the models in Experiment 3 with the chosen baseline benchmarks. *(B)* denotes a baseline model while (**i*) denotes the best performing model from the previous i^{th} experiment.

Model	ΔSSIM (↑)	ΔFID (↓)	ΔLPIPS (↓)
	$SSIM_0$ = 0.463	FID_0 = 6.197	$LPIPS_0$ = 0.271
Model 6 (*3)	0.075	-4.640	-0.029

Table 10. Improvement in metrics for each of Experiment 3's model variation in comparison to the best performing model variation from Experiment 2 (e.g., $SSIM_i - SSIM_0$).

Fig 23. Results obtained from the model variation in Experiment 3. Model 5(*2) refers to the best performing model observed in Experiment 2.

As seen in the first two samples, the long-standing problem that occluded fashion attributes present in the style image are not well represented in the generated image has been solved to a limited extent. Nevertheless, the inclusion of the Global Styles Fusion Module helped in the replication of styles by embedding them at different spatial resolutions. These improvements are exemplified in the remaining samples of Fig 23.

Observing the first two samples, the model still struggles with replicating styles at small and precise locations. For example, the model fails to replicate the white ankle long loose-fitting pants the female subject is wearing. This led to poorer performance in samples such as the first two columns. Meanwhile, the model is better able to learn the styles embedding at

different spatial resolutions due to the novel architecture of generating style embeddings from feature maps of different dimensions generated by the Swin Transformer. In this light, the model can learn the global features of the style present, such as colour, while also learning more local features of the style present, such as interweaving different layers of clothing coherently. This provides additional advantages as compared to passing in a global representation of the styles since colours from different fashion attributes, especially those in smaller pixel neighborhoods, may be infused altogether causing inaccurate colours.

While the closer replication of styles resulted in improvements in the reported metrics, another contributing factor was the generation of plain and simpler backgrounds which mirrored those seen in samples from the DeepFashion MultiModal dataset. This is because the same human segmentor from the H_E module was used to mask out the human in the style image before it was passed to the Style Extractor Module H_S. Hence, bits of the background were also used as conditioning in the cross-attention layers in the decoder.

For further improvements, a higher *ip_rank* could be used such that the style embeddings at each spatial resolution are represented with more tokens. This could lead to further alignment between the generated styles and the styles present in the style image. Alternatively, the projection layers used to transform the style embeddings from the K transformer blocks output could be further improved upon.

4.3.5 Experiment 4 - Fine-tuning of the Decoder for smoother backgrounds

Experiment 4 intends to resolve the challenge of generating a homogeneous plain background that is cohesive with the generated foreground, as seen by the distinct patches of colour differentiating the foreground and the background in the generated output of Model 5 from Experiment 2.

To resolve this, LoRA fine-tuning is adopted by introducing additional fine-tuning parameters for the query Q and value V vectors of the cross-attention layers. Fine-tuning was performed with a constant text prompt of "a person. plain studio background" and this allows for the model to learn how to better align the noisy latent in the Stable Diffusion backbone with the text conditioning of a simple studio background.

It is important to note that this experiment was performed without the Global Styles Fusion Module so that any improvements in results can be appropriately attributed to the proposed fine-tuning approach. The remaining training hyperparameters, choice of hyperparameter for H_E and H_M remains the same as in Experiment 2.

LoRA Fine-tuning	lora_q_rank	32
	lora_q_scale	1.0
	lora_v_rank	32
	lora_v_scale	1.0
Train Related	Number of training steps	82,000
	Training Dataset size	11,000 samples
	Resumed Checkpoint	expt2-model5-gs155k.ckpt

Table 11. Hyperparameters used for LoRA fine-tuning. expt2-model5-gs155k.ckpt refers to the pre-trained checkpoint from Model 5 of Experiment 2, up till a global step of 155,000.

Fig 24. Comparison of results obtained from the best model variations from Experiment 2, 3 and 4.

Model	SSIM (↑)	FID (↓)	LPIPS (↓)
PIDM *(B)*	**0.656**	0.390	0.184
CFLD *(B)*	0.660	**0.195**	**0.183**
Naïve HF implementation *(B)*	0.566	7.250	0.234
Baseline Visconet *(B)*	0.401	9.790	0.306
Model 5 (*2)	0.463	6.197	0.271
Model 6 (*3)	**0.538**	1.557	**0.242**
Model 7 (*4)	0.529	**1.407**	0.243

Table 12. Results obtained from the model variation in Experiment 3. Model 5(*2) refers to the best performing model observed in Experiment 2.

Model	ΔSSIM (↑)	ΔFID (↓)	ΔLPIPS (↓)
	$SSIM_0$ = 0.538	FID_0 = *1.557*	$LPIPS_0$ = 0.242
Model 6 (*3)	-0.009	-0.150	0.001

Table 13. Improvement in metrics for each of Experiment 3's model variation in comparison to the best performing model variation from Experiment 2 (e.g., $SSIM_i - SSIM_0$).

From Fig 24 and Table 12, it is evident that the fine-tuning has indeed provided closer alignment between the noisy latent in the denoising U-Net and the conditioning text prompts. While the fine-tuning was not able to solve the issue of replicating the styles present in the source image to its fullest, it was able to solve the intended problem of generating homogeneous plain backgrounds that fused both the foreground and background.

With the fine-tuning, the model is consistently able to generate images with plain studio backgrounds when given the prompt of "a person. plain studio background.". This consequently resulted in better metrics as the generated images are now of greater similarity to the images present in the DeepFashion MultiModal dataset. Even then, how closely aligned the generated images are to those in the DeepFashion MultiModal dataset, characterised by plain studio backgrounds, could be fairly controlled using the scaling alpha α parameters used in the LoRA matrices. Fig 25 demonstrates a few examples where the α parameters were varied for several use cases of generating a diverse range of illustrative backgrounds. For brevity, results showing only variations in the α parameter for the fine-tuned V matrix is shown.

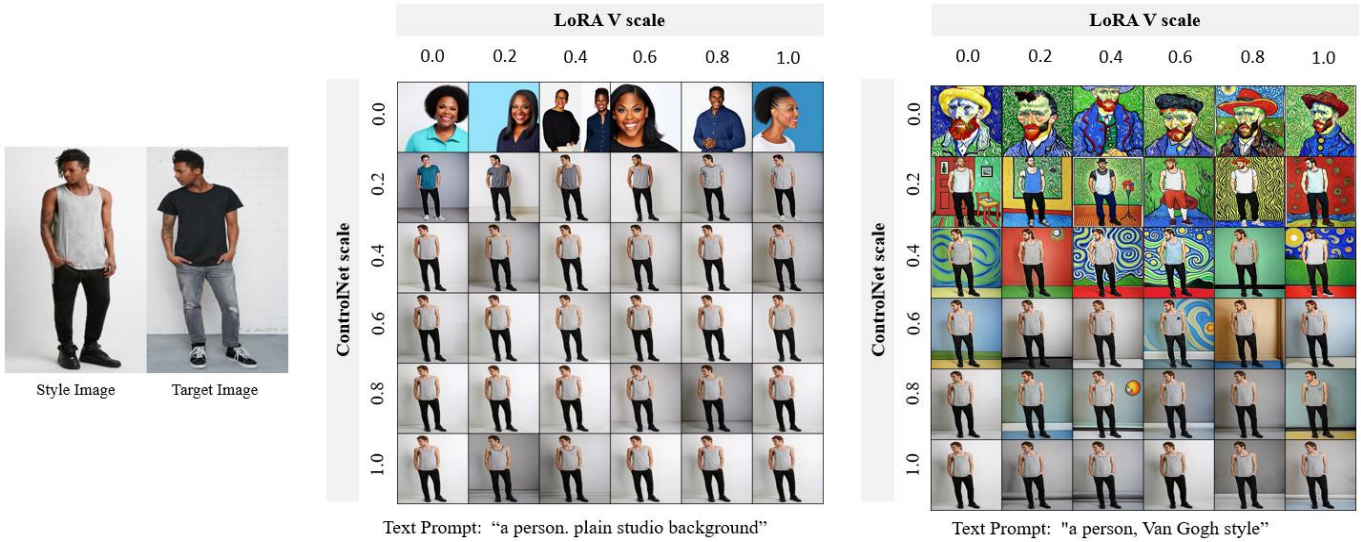


Fig 25. Comparison of results obtained from different scale values used for the ControlNet and the LoRA V matrix scale for different text prompts used. The text prompts used focused on the describing the styles of the background.

From Fig 25, the pre-trained Stable Diffusion backbone has still managed to retain its generative capabilities by demonstrating reasonable generative performance for a diverse range of background styles ranging from plain simple styles (on the left of Fig 25) to more complicated styles (on the right of Fig 25).

The styles in the DeepFashion MultiModal (DFMM) dataset are characterised by their simplicity and plainness. As the ControlNet scale increases, the generation tends to align more with the styles observed in the DFMM dataset. Meanwhile, as the LoRA V scale increases, simpler and more uniform backgrounds are being generated albeit there is still the presence of artifacts. This aligns with the constant text prompt used during the fine-tuning process which trains the model to generate simple and plain backgrounds like those present in the DFMM dataset.

In summary, these results showed that the approach of fine-tuning the Q and V matrices in the cross-attention layers of the decoder's attention blocks resulted in a huge increase in performance for the task of Human Pose Transfer with the DFMM dataset. This is because simpler backgrounds with fewer artifacts were generated after fine-tuning. Meanwhile, the model's original generative capability is retained and a balance between the original generative capabilities and the effects of the LoRA fine-tuning can be achieved by tuning the ControlNet scale values and the LoRA scale values.

Comparing the metrics obtained for Models 6 and 7, ensuring the generation of a simpler and plain background was the largest contributing factor to improving especially the network-based metrics that relied on feature maps to determine how identical the distributions of the generated and ground truth images were.

Nonetheless, it is worth noting that direct comparisons of the generated samples from Models 6 and 7 demonstrate the effectiveness of the Global Style Fusion Module in replicating the styles present in the style image. This could in turn lead to improvements in non-network-based metrics that measure how similar two images are in terms of qualities such as brightness, contrast, and similarity. Hence, it would be worth considering to further train an architecture that incorporates both the Global Style Fusion Module and LoRA fine-tuning of the decoder.

4.4 Ablation Studies

While the previous experiments demonstrate the usefulness of each module added sequentially, there still lacks a fair analysis of the advantages of adding each module piecewise to the overall architecture. This is because in the previous experiments, parameters such as the dataset size which is crucial in preventing overfitting may not have remained constant due to long training time involved. To better analyse the effectiveness of each module, an ablation study was conducted as well. Each ablation study involved the removal of a module to allow for the effectiveness of each module to be studied in isolation.

Due to time constraints, each ablation's training was stopped early when no more significant improvements were observed.

Ablation	**No. of training steps ('000s)**	**ControlNet Module (Pose only)**	**ControlNet Module (Pose with H_E)**	**Global Styles Fusion Module, H_S and H_A**	**LoRA Fine-tuning**
1	55	Yes	No	No	No
2	155	Yes	Yes	No	No
3	100	Yes	Yes	Yes	No
4	25	Yes	Yes	Yes	Yes

Table 14. Modules and number of training steps used in each ablation study.

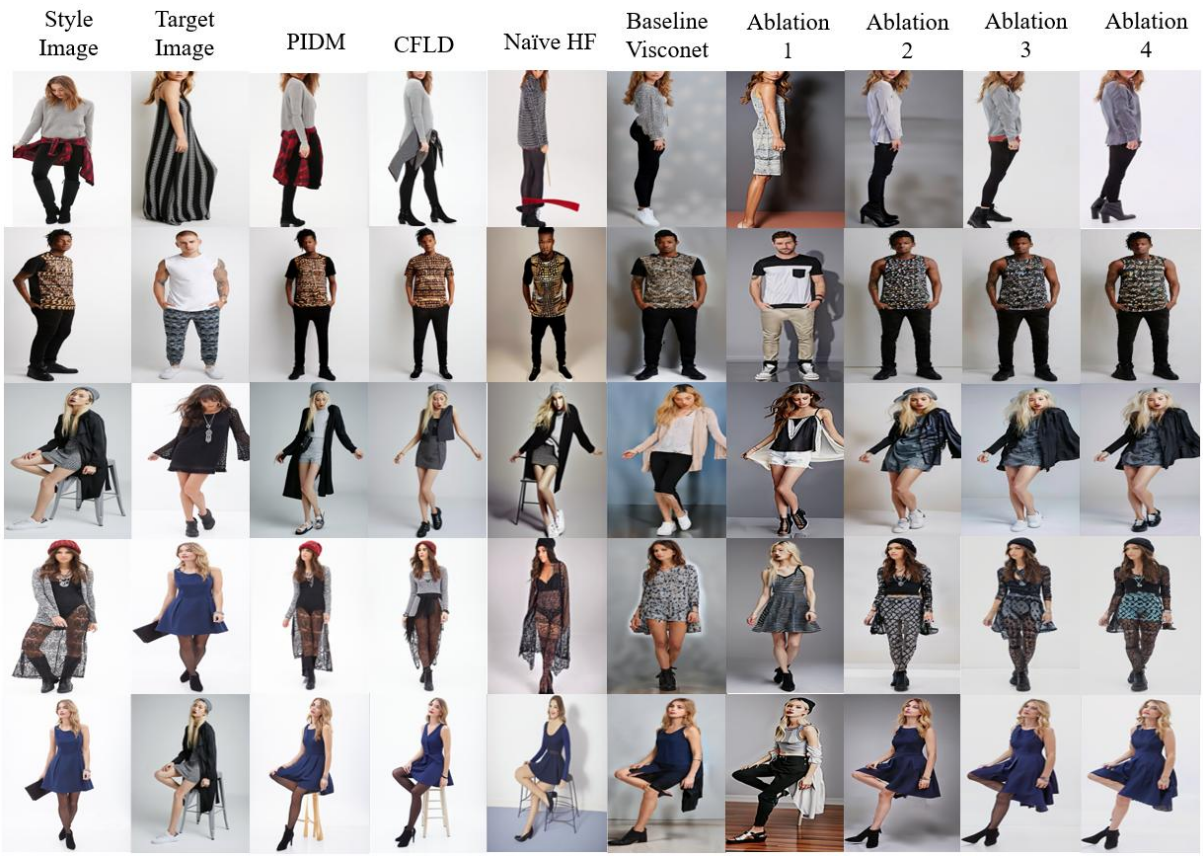


Fig 26. Comparison of the results obtained from the baselines and each ablation's model.

Model	SSIM (↑)	FID (↓)	LPIPS (↓)
PIDM *(B)*	**0.656**	0.390	0.184
CFLD *(B)*	0.660	**0.195**	**0.183**
Naïve HF implementation *(B)*	0.566	7.250	0.234
Baseline Visconet *(B)*	0.401	9.790	0.306
Ablation 1	0.407	8.004	0.308
Ablation 2	0.467	6.566	0.266
Ablation 3	0.538	**1.557**	**0.242**
Ablation 4	**0.541**	1.604	0.243

Table 15. Metrics obtained using the model from each ablation.

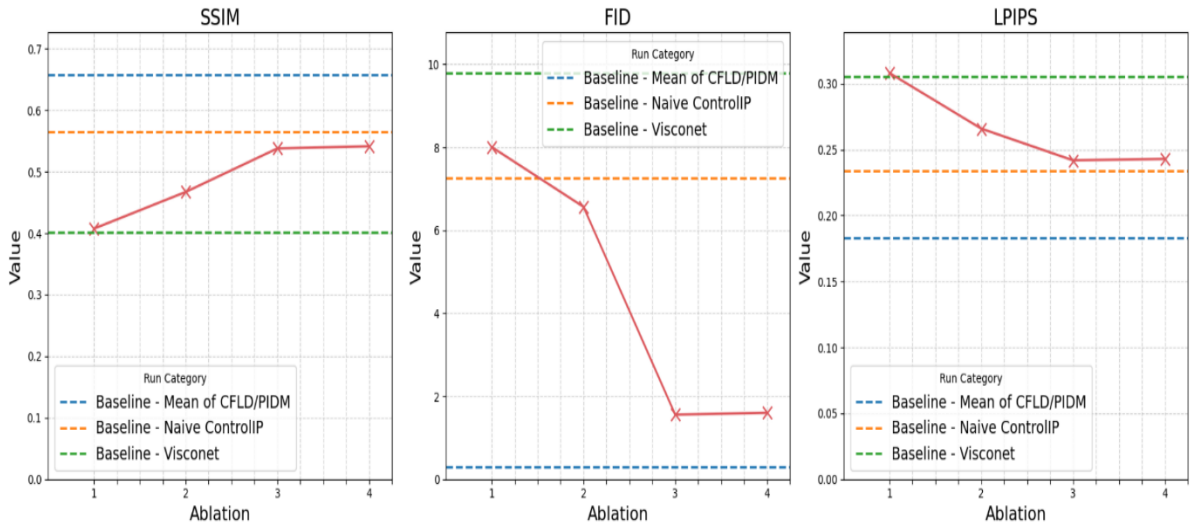


Fig 27. Graphical visualisation of the key metrics obtained from each ablation in comparison to the baselines.

By comparing the results from each ablation with the previous, the effectiveness of each module can be observed.

In Ablation 1, only pose information was used as input via the ControlNet Module while no conditioning about the fashion attributes were given to the Style Encoder Module, H_E in the ControlNet Module. While there was strict adherence to the conditioning pose, the generated fashion attributes remained largely random. This suggests the effectiveness of passing in pose information into the ControlNet to perform pose transfer.

In Ablation 2, segmented fashion attributes from a pre-trained fashion segmentor module [24] was used to extract fashion attributes present in the style image. These images were then given to the Style Extractor Module H_E and was used to guide the generation in areas defined by the Control Feature Mask Module H_M. While there is room for further improvements, the improvement in performance seen in Ablation 2 suggests the effectiveness of passing in pose and style information into the modified ControlNet module proposed in this work to ensure that styles present in the style image are replicated and not replicated indiscriminately.

In Ablation 3, the masked out human from the style image is first encoded using the same pre-trained VAE encoder as the Stable Diffusion backbone. Embeddings at different spatial resolutions are then generated using the proposed Style Extractor Module H_S and are passed into the denoising U-Net decoder via the cross-attention layers using the proposed Attention Fusion Module H_A. While Ablation 2 demonstrated the model's ability to replicate styles present in the style image, small minor details which may not be represented well by H_E via the ControlNet module tend to be omitted in the generated images. One such example is the

absence of the black legging the female subject is wearing, which was omitted in Ablation 1 and 2, but finally picked up in Ablation 3.

In Ablation 4, additional LoRA matrices were added to the Q and V matrix of the cross-attention layers in the denoising U-Net decoder while freezing all other weights in the Stable Diffusion backbone, ControlNet Module and Global Styles Fusion Module. The aim was to enable the model to learn how to generate smoother backgrounds with less artifacts by introducing additional weights that helps the model to align the noisy latent, passed in via the query projection matrix, with the text conditioning, passed in via the key and value projection matrix, which described a plain and simple background. While this helped in generating smoother and plainer background in some samples, the model from Ablation 4 seems to exhibit poorer performance as seen by the generated baseline samples and metrics. A possible reason could be that more steps are required to update the weights, or the Global Style Fusion Module's weights should be updated simultaneously as both modules affect the output values of the attention blocks in the decoder.

4.5 Final Proposed Model

From the ablation studies, the effectiveness of both the ControlNet Module and Global Styles Fusion Module have been demonstrated. Meanwhile, the effectiveness of fine-tuning the denoising U-Net decoder requires longer training time and more experimentation which was unfortunately not possible due to the time constraints of this project.

Module	**Component**	**Key Parameters**	**Trainable Params.**
ControlNet	ControlNet branch	context_dim: 1024	364.2M
	H_E (*Resampler*)	image-encoder = DINOv2, depth = 8, num-queries = 16, embedding-dim = 1024	102.8M
	H_M	dilate-kernel-size = 5, dilate-iterations = 8, max-distance = 30, distance-scale = 'exp', distance-scale-factor = 3	0
Global Styles Fusion	H_S	transformer-block-dim = 256, transformer-depth = 1, get-featuremaps-method = 'attn', proj-out-dim = 77	62.6M
	H_A	ip-embed-seq-len = 77, ip-rank = 4	23.4M
LoRA *(optional)*	LoRA	lora-rank = 32, lora-q-scale = 1.0, lora-v-scale = 1.0	0.9M

Method	**Pose Info. & Annotation**	**Trainable Params.**
PIDM	2D OpenPose of 20 channels (Pose RGB Image & 17 joint heatmaps)	688.0M
CFLD	2D OpenPose of 20 channels (Pose RGB Image & 17 joint heatmaps)	248.2M
ViscoNet	2D OpenPose of 3 channels (Pose RGB Image)	364.2M
ViscoNet Improved (Ours)	2D OpenPose of 3 channels (Pose RGB Image)	553.9M

Table 16. Summary of the number of parameters and pose information for each module in the final chosen model.

4.6 Advantages from existing baselines

4.6.1 Harmonisation of Text Prompt

One of the biggest advantages of the proposed approach is the capability to incorporate text prompts to augment the generation. In the context of Pose Transfer, text prompts could be used to reinforce the stylistic conditioning of fashion attributes by passing in prompts that describe the fashion attributes present. Alternatively, text prompts could also be used to provide additional details regarding the pose to make the generation more cohesive. Finally, leveraging on the generative capabilities of the pre-trained Stable Diffusion backbone, Pose Transfer could be performed with a variety of backgrounds.

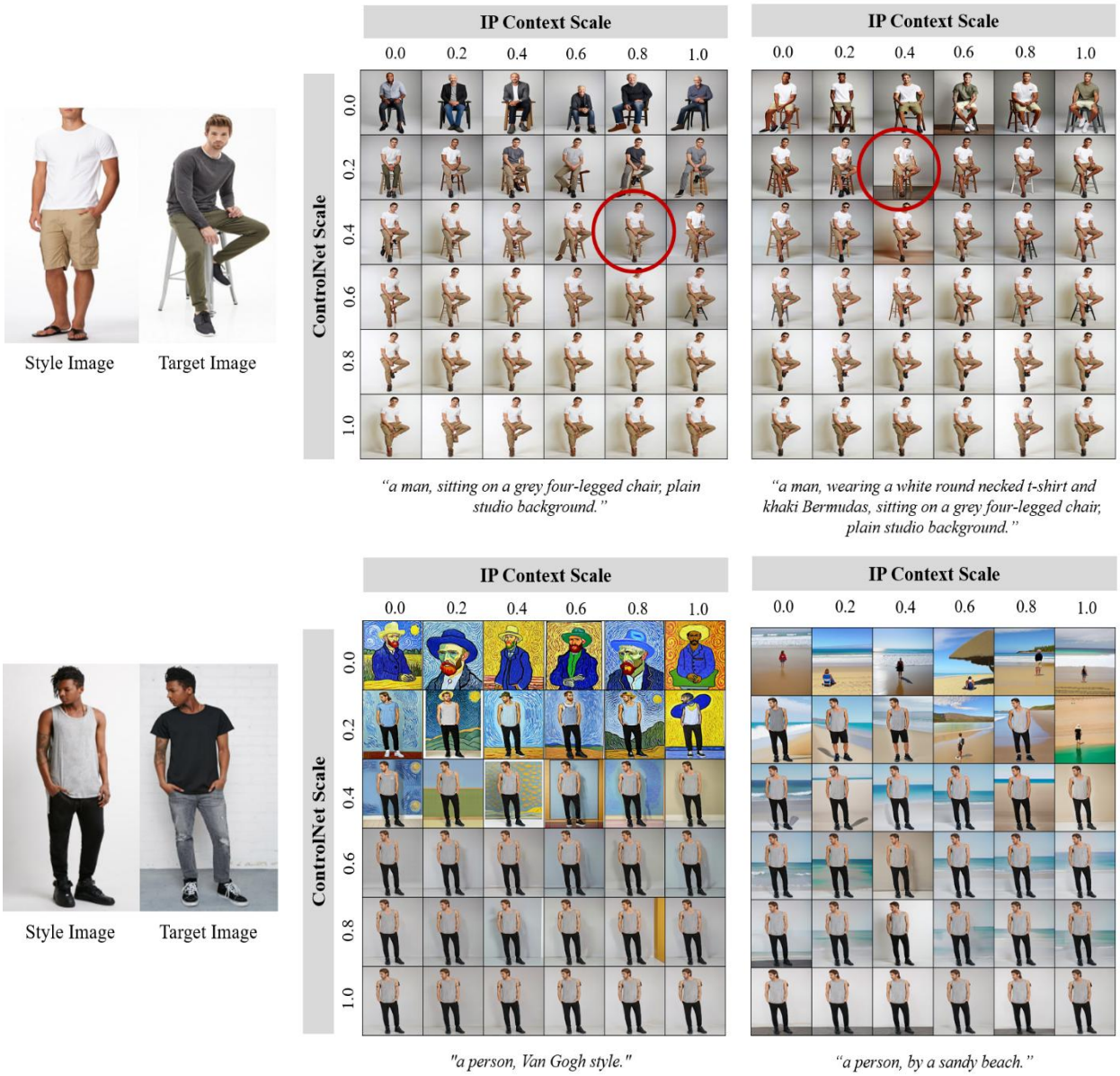


Fig 28. Examples of using text prompt to either include details about the pose (top left), describe the fashion attribute present (top right), or details about the overall style or background (bottom left and right) to alter the generation outputs.

Moreover, to ensure an effective harmonisation of conditionings from different modalities, different scale values can be applied to the conditionings from the ControlNet and Global Styles Fusion Module.

Observing the generated outputs from the image grid along the same column, increasing the value of the ControlNet scale tends to result in generated images looking more like images from the DeepFashion MultiModal dataset. Meanwhile, observing generations across the same row, there is not much obvious change in generations as the IP Context Scale is increased, which could be due to the low value of the *ip_rank* being used. Nevertheless, by appropriately selecting a combination of scale values, the conditioning from the text prompt could be effectively integrated into the generation to generate cohesive images that best represent the fashion attributes present in the style image and the pose in the target image.

4.6.2 Generalised Pose Transfer

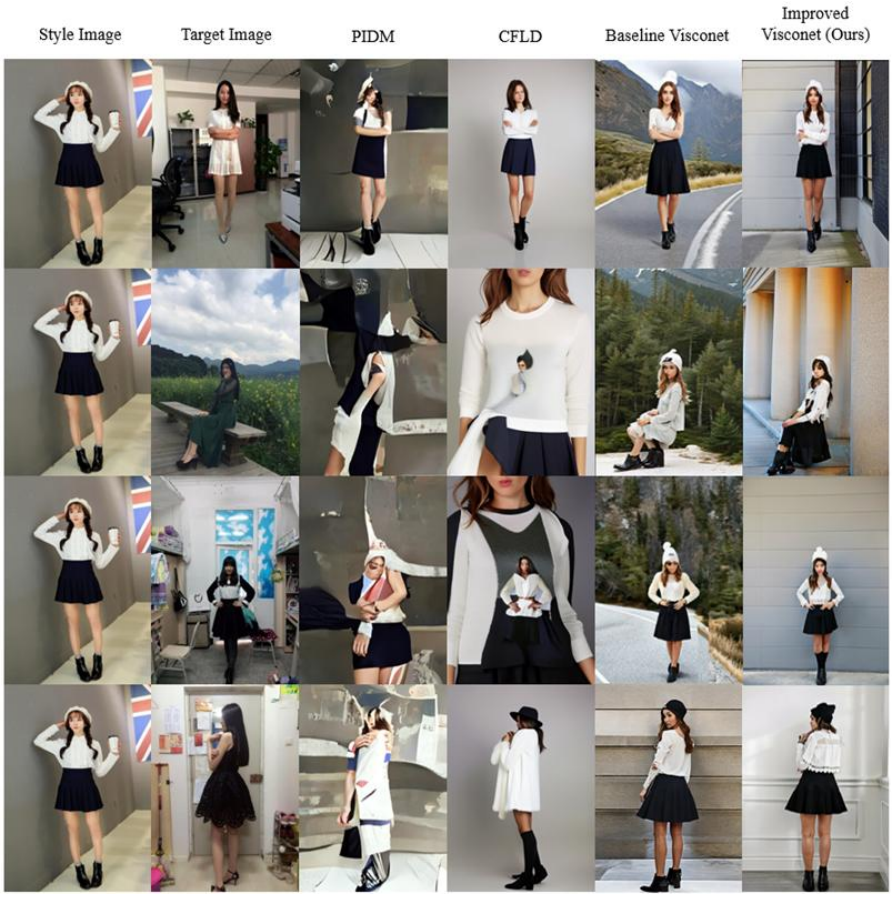
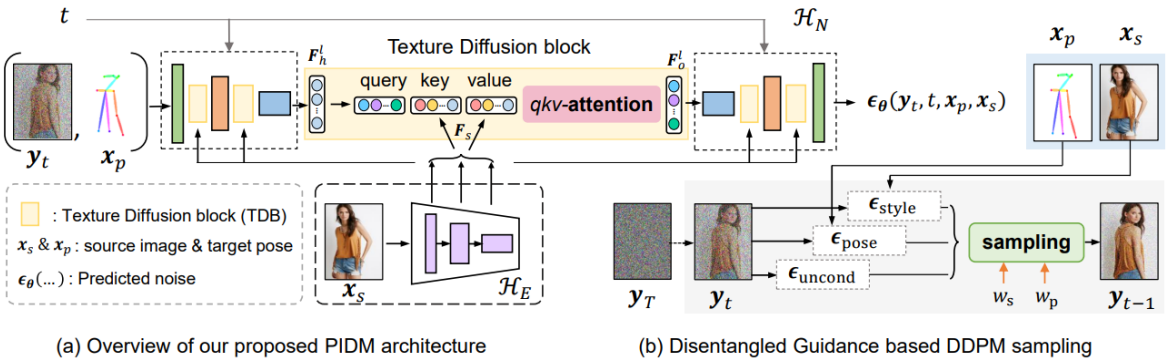


Fig 29. Generated samples from the DeepFashion2 dataset [42], which is an extension of the DeepFashion dataset consisting of images with day-to-day backgrounds and styles.

Fig 29 demonstrates the model's ability to perform Pose Transfer independently from the background. While the generated outcomes from PIDM and CFLD tend to be noisy without generating the subject in the desired target pose, both the baseline and our improved Visconet generate the subject in the desired target pose while replicating the styles present in the style image.

While the generation qualities for both the baseline and our improved Visconet can be further improved, there are still distinct improvements when performing Pose Transfer using in-the-wild samples. These "in-the-wild" samples tend to be images from everyday life with varying backgrounds and human-to-image ratios. The human-to-image ratio refers to the proportion of the human subject relative to the entire frame. This differs from those observed in the DeepFashion MultiModal dataset which largely consists of high-definition images with plain and simple backgrounds and the human subject centred in the image.

The difference in performance could be explained by the difference in Visconet's architecture and the chosen baseline models, PIDM and CFLD.



(a) Overview of our proposed PIDM architecture (b) Disentangled Guidance based DDPM sampling

Fig 30. Architecture diagram of one of the chosen baselines, PIDM. Taken from [14].

In PIDM, the style image is first encoded using an encoder, and multi-scale features are derived by taking the output from the different layers of H_E. These multi-scale features are then used as the keys K and values V in the cross-attention layers in both the encoder and the decoder. CFLD adopts a similar concept but further refines the conditioning from the style images. This is achieved by using a hierarchical vision transformer to get the multi-scale features, also referred to as coarse features, and learning a set of learnable queries to represent the style image using the final scaled features, also referred to as fine features. The fine features are then injected as keys K and values V in both the encoder and the decoder while the coarse features are only injected as bias along with the queries Q in the decoder.

These modifications allow for the styles present in the style image to be replicated while also ensuring a closer alignment between the distribution of the generated images and the distribution of images the model was trained on. As network-based metrics such as FID and LPIPS measure how similar the distribution of generated images to the original image's distribution, this allows the baseline models PIDM and CFLD to achieve better metrics when tasked to generate samples from the same distribution it was trained on, namely the DeepFashion MultiModal dataset in this work. However, as the model learns to approximate the distribution of the image dataset it was trained on better, it loses its generalisation capabilities and may generate incoherent samples when generating "in-the-wild" samples. On the other hand, the improved Visconet proposed in this work addresses this issue and is not constrained to predicting within the DeepFashion multimodal dataset distribution.

In the improved Visconet proposed in this work, conditionings from the style image are only injected in the decoder through a conditioning mask. While this means that the generated outputs may not be as closely aligned, it allows the model to preserve its generalisation capabilities by not encoding the style conditionings into the same latent space. Introducing style conditionings in the decoder only also allows the model to better attend to the stylistic nuances of each sample thereby leading to a more diverse range of outputs. Hence, while the improved Visconet proposed in this work may perform slightly worse on the DeepFashion MultiModal dataset, it still can have decent performance on the "in-the-wild dataset".

5 Conclusion and Future Work

The aim of this project was to implement a novel model architecture that incorporated the latest image prompt adapters with pre-trained text-to-image models to perform Human Pose Transfer using multi-modality prompts. While there exist many previous works for the task of Human Pose Transfer, these works mainly leverage image conditioning only. The advent of image prompt adapters allowed for the effective use of image conditioning and text conditioning. Coupled with the impressive generative capabilities of large pre-trained text-to-image models like Stable Diffusion in generating a diverse range of images, the proposed model architecture aimed to incorporate both text and image conditioning to perform Human Pose Transfer over a diverse range of samples.

Moreover, existing works tend to lack the ability to generalise. This is due to architectural differences that let these model learns the distribution it was trained on more closely at the expense of being able to generalise over a wider range of samples. With the novel architecture proposed in this work, the proposed model can achieve decent performance on a wide range of samples even though it was only trained on the DeepFashion MultiModal dataset. Even though the proposed model performs slightly worse in the DeepFashion MultiModal dataset, it retained its ability to generalise to samples outside its training dataset. This is due to the phenomenon where the model tends to exhibit poorer performance in other tasks when it is trained to fit another task better.

Currently, performance on the DeepFashion MultiModal dataset could be further improved. A possible approach would be to use higher rank matrices to decompose and reconstruct the keys K and values V matrices used to represent the style conditionings in the decoder, but care should be taken in ensuring that the model does not overfit the training dataset.

Alternatively, the improved proposed Visconet model could be trained using a more diverse dataset consisting of samples from the more common everyday images from more datasets. This is because most works currently are still limited to the DeepFashion dataset and further research can be performed to achieve Pose Transfer across a more diverse range of samples.

6 References

[1] Z. Jia, Z. Zhang, L. Wang and T. Tan, “Human Image Generation: A Comprehensive Survey,” *ACM Computing Surveys,* vol. 56, no. 11, pp. 1-39, 2018.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. & Ommer, “High-resolution image synthesis with latent diffusion models.,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022.

[3] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti and J. ... & Jitsev, “Laion-5b: An open large-scale dataset for training next generation image-text models.,” in *Advances in Neural Information Processing Systems*, 2022.

[4] 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.*, 2023.

[5] H. Ye, J. Zhang, S. Liu, X. Han and W. Yang, “IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models.,” arXiv, 2023.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair and Y. ... Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems,* vol. 27, 2014.

[7] D. P. Kingma, “Auto-encoding variational bayes,” arXiv, 2013.

[8] T. Karras, S. Laine and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[9] P. Esser, E. Sutter and B. Ommer, “A variational u-net for conditional appearance and shape generation.,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2018.

[10] K. Sarkar, L. Liu, V. Golyanik and C. Theobalt, “Humangan: A generative model of human images,” in *2021 International Conference on 3D Vision (3DV)*, 2021.

[11] J. Ho, A. Jain and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems,* vol. 33, pp. 6840-6851, 2020.

[12] K. Zhang, M. Sun, J. Sun, B. Zhao, K. Zhang, Z. Sun and T. Tan, “Humandiffusion: A coarse-to-fine alignment diffusion framework for controllable text-driven person image generation.,” 2022.

[13] S. Y. Cheong, A. Mustafa and A. Gilbert, “Upgpt: Universal diffusion model for person image generation, editing and pose transfer.,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[14] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah and F. S. Khan, “Person image synthesis via denoising diffusion model.,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[15] Y. Lu, M. Zhang, A. J. Ma, X. Xie and J. Lai, “Coarse-to-fine latent diffusion for pose-guided person image synthesis.,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2024.

[16] S. Y. Cheong, A. Mustafa and A. Gilbert, “Visconet: Bridging and harmonizing visual and textual conditioning for controlnet.,” arXiv, 2023.

[17] O. Ronneberger, P. Fischer and a. T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, 2015.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal and e. al., “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021.

[19] A. Vaswani, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[20] OpenAI, “Hugging Face,” 2021. [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14. [Accessed 20 1 2025].

[21] Z. Cao, T. Simon, S. E. Wei and a. Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] R. A. Güler, N. Neverova and a. I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Y. Choi, S. Kwak, K. Lee, H. Choi and a. J. Shin, “Improving diffusion models for virtual try-on,” 2024.

[24] M. Djaga, “SegFormer B2 Clothes,” 2021. [Online]. Available: https://huggingface.co/mattmdjaga/segformer_b2_clothes. [Accessed 20 1 2025].

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[26] benjamin-paine/stable-diffusion-v1-5, “Hugging Face,” 2024. [Online]. Available: https://huggingface.co/benjamin-paine/stable-diffusion-v1-5. [Accessed 22 1 2025].

[27] lllyasviel/control_v11p_sd15_openpose, “Hugging Face,” [Online]. Available: https://huggingface.co/lllyasviel/control_v11p_sd15_openpose. [Accessed 22 1 2025].

[28] H94/IP-Adapter, “Hugging Face,” [Online]. Available: https://huggingface.co/h94/IP-Adapter. [Accessed 22 1 2025].

[29] L. AI, “Structural similarity index measure (SSIM) — TorchMetrics 1.3.1 documentation,” 12 3 2025. [Online]. Available: https://lightning.ai/docs/torchmetrics/stable/image/structural_similarity.html.

[30] L. AI, “Frechet inception distance (FID) — TorchMetrics 1.3.1 documentation,” Lightning AI, 12 3 2025. [Online]. Available: https://lightning.ai/docs/torchmetrics/stable/image/frechet_inception_distance.html.

[31] L. AI, “Learned perceptual image patch similarity (LPIPS) — TorchMetrics 1.3.1 documentation,” Lightning AI, 12 3 2025. [Online]. Available: https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity.html.

[32] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang and a. X. Bai, “Progressive pose attention transfer for person image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy and a. Z. Liu, “Text2Human: Text-driven controllable human image generation,” *ACM Transactions on Graphics (TOG),* vol. 41, no. 4, p. 1–11, 2022.

[34] S. AI, “Stability AI/Stable Diffusion,” [Online]. Available: https://github.com/Stability-AI/stablediffusion. [Accessed 22 1 2025].

[35] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov and .. a. P. Bojanowski, “DinoV2: Learning robust visual features without supervision,” arXiv, 2023.

[36] L. AI, “GPU Intermediate: PyTorch Lightning Documentation,” [Online]. Available: https://lightning.ai/docs/pytorch/stable/accelerators/gpu_intermediate.html. [Accessed 22 1 2025].

[37] C. Liu, X. Yang, L. Wang and a. J. Tang, “In-Shop Clothes Retrieval with Deep Fashion,” [Online]. Available: https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion/InShopRetrieval.html. [Accessed 22 1 2025].

[38] X. Zheng, Y. Yang and a. A. Hauptmann, “Scalable person re-identification: A benchmark,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[39] PyTorch, “DeepLabV3+ (ResNet-101) model,” [Online]. Available: https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/. [Accessed 22 1 2025].

[40] J. Y. Peh, “Visconet configuration files,” [Online]. Available: https://github.com/jinyangp/visconet/tree/main/configs. [Accessed 22 1 2025].

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[42] Y. Ge, R. Zhang, X. Wang, X. Tang and P. Luo, “DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] Y. Ren, X. Fan, G. Li, S. Liu and a. T. H. Li, “Neural Texture Extraction and Distribution for Controllable Person Image Synthesis,” arXiv.org, 2022.