GraVITON: Graph based garment warping with attention guided inversion for Virtual-tryon

Sanhita Pathak[1], Vinay Kaushik[2], and Brejesh Lall[1]

[1] Indian Institute of Technology, New Delhi, India
Sanhita.Pathak@dbst.iitd.ac.in, brejesh@ee.iitd.ac.in
[2] Indian Institute of Information Technology Sonepat, Haryana, India
vkaushik@iiitsonepat.ac.in

Abstract. Virtual try-on, a rapidly evolving field in computer vision, is transforming e-commerce by improving customer experiences through precise garment warping and seamless integration onto the human body. Existing methods such as TPS and flow address the garment warping, but overlook the finer contextual details. In this paper, we introduce a novel graph based warping technique which emphasizes the value of context in garment flow. Our graph based warping module generates warped garment as well as a coarse person image, which is utilised by a simple refinement network to give a coarse virtual tryon image. We then exploit a latent diffusion model to generate the final tryon, treating garment transfer as an inpainting task. The diffusion model incorporates a Decoupled Garment Attention Adaptor(DGAA) for attention based diffusion inversion of visual and textual information. Our method, validated on VITON-HD and Dresscode datasets, showcases substantial state-of-the-art qualitative and quantitative results showing considerable improvement in garment warping, texture preservation, and overall realism.

Keywords: Virtual tryon · Optical Flow · Graph · Latent Diffusion models

1 Introduction

With the evolving shopping trends, ecommerce platforms have started catering to the customer needs keeping in sync with the emerging requirements. In the apparel industry, this has come into view as virtual tryon, which can provide a real inshop experience to the customers. The image based tryon methods [3, 12] have proven to be more practical when compared to the 3D [15] models which require modelling of the person for a realistic tryon synthesis which is quite labor-some.

To produce a perfect tryon result, the person and garment variability has to be prioritised while formulating the tryon pipeline. Although various studies have synthesized compelling results on the benchmarks [21] [6] [11], there still exist some paucity in terms of realism.

The tryon technique was first introduced by VITON [11], which used TPS warping for solving the problem of warping garments in virtual tryon. CPV-TON [25] preserved texture, but lacked perfect alignment, while the flow based approaches [3, 16, 33] learnt robust structural alignment but lacked texture consistency. Other methods [20, 26] focused more on improving the generation by using various synthesis models such as GANs and recently diffusion [14, 20, 28]. Amid all the advancements in various stages of virtual tryon, there are still considerable gaps such as learning better garment warp, handling occlusion, pose transformations, generating consistent texture, etc. present, that leave a great scope of improvement.

The current methods [3, 33] typically model the flow as result of correlations (utilising either a simple convolution network or feature correlation) between features across garment and reference images(pose,agnostic). These approaches mainly encode the point wise correspondence between an image feature pair(s) while neglecting the intra-relations among pixels within regions [19]. There's a need to capture discriminative features for region and shape representations. Thus, decoupling the garment context from the warping procedure, and simultaneously transferring the region and shape prior of garment context to warping network can aid in learning an optimal garment warp.

Motivated from AGFLOW [19], which introduces iterative graph based flow estimation, we propose a solution to the aforementioned problem on warping by building a novel graph based garment warping module, which embeds context into learning garment warp onto the warping pipeline. The proposed Graph based flow warping module (GFW) learns to match features conditioned on garment context. This allows object's spatial neighbourhood to be well aggregated and thus largely decreases the uncertainty of ambiguous warping of garment.

Diffusion models [7] currently stand as the top-performing models; when compared to the flow and TPS based counterparts [2, 25, 33]. However, maintaining texture consistency during warping poses a challenge. Recent diffusion-based approaches, exemplified by LaDI-VTON [20], StableViton [14], dci-vton [8], CAT-DM [32] address this challenge by leveraging textual and visual context for virtual try-on generation, treating it as a conditional image inpainting task. To achieve this, LaDI-VTON [20] proposes an inversion module, where image features are extracted from an image encoder and mapped to new word embeddings by a trainable network and then concatenated with text embeddings. StableV-TON [14] utilises a ControlNet model that is directly conditioned on straight garment, incorporating a zero-conv cross attention block. CAT-DM [32] initiates a reverse denoising process, utilising an LDM, with an implicit distribution generated by a pre-trained GAN-based model, thereby reducing the sampling steps without compromising generation quality. In the cross-attention module of LaDI-VTON [20], merging straight cloth features and text features into the cross-attention layer only accomplishes the alignment of image features to text features, and potentially misses some image-specific information and eventually leads to only coarse-grained controllable generation with the reference image. This leads to texture transfer artefacts in some scenarios. For a better tryon

inversion, we propose Decoupled Garment Attention Adaptor(DGAA), which adds an additional cross-attention layer only for image features [31].

The contributions of our proposed work are as follows:

- We introduce a Graph based flow warping module(GFW), that guides the appearance flow by providing garment pixel neighbourhood context into flow prediction.
- We propose a Decoupled Garment Attention Adaptor (DGAA), enriching latent space diffusion inversion for a realistic tryon.
- Extensive experimentation and rigorous validation demonstrates that our method achieves state-of-the-art performance compared to existing prominent methods.

2 Related Works

2.1 Virtual tryon

Given a set of straight cloth and a person image, the goal of virtual tryon is to seamlessly warp the garment and overlay it onto the target person image. The initial work that introduced the garment warping and a generated complete person tryon was VITON [11]. Other methods [10, 12, 25, 30, 33] followed a similar two stage warping and generation pipeline, which learnt TPS or affine transformation parameters for computing garment warp. Although TPS preserves the texture of warped garment better than to that of it's flow based counterparts, incorporating flow achieves optimal garment alignment with the changing human pose. In order to achieve the realism in the final tryon, it is crucial to formulate a robust garment deformation module. This is usually achieved by the deformation of control points with an energy function (radial basis function) in TPS based pipelines (Thin Plate Spline) [25], and by computing per pixel appearance flow followed by target view synthesis in flow based pipelines. The flow based warping learns dense per-pixel correspondence [3, 12, 16, 33], when compared to the TPS based methods which are unable to capture such local warp details.

2.2 Graph neural networks in flow

Optical flow is the task of estimating dense per-pixel correspondence between images. GMFlow [27] introduced vision transformers for computing optical flow, but its heavy computational dependencies made it less diversely applicable. AGFlow [19] exploited the scene/context information, utilising graph convolutional networks, and incorporated it in the matching procedure to robustly compute optical flow. Virtual tryon entails computation of appearance flow [10, 33], to warp the source cloth based on the reference person features (pose, densepose, etc.) GPVTON [33] tried to address the local deformations by applying a part wise flow based deformation, where the garment is disintegrated and deformed separately into three regions, one for each upper body part. GPVTON is not able to jointly optimise the local and global deformations. Another work KGI [17]

utilised graph to predict the garment pose points guided by human pose which inpainted the predicted region using human segmentation. The method failed to achieve the precision in tryon alignment due to sparse guiding points to guide the dense pixel warping for garment texture unlike in flow methods. Hence, motivated by AGFlow [19], in this work we have shown that GCNs can help the garment warping by focusing on the pixel level deformations establishing a dense correlation that helps in preserving the local details post deformation, which is ideally faced by all the flow based garment warping methods.

2.3 Diffusion Models

Diffusion models marked research has become a foundational area in the field of image synthesis [7] because of its high quality image generation. Tasks such as image-to-image translation [24], image editing [1], text-to-image synthesis [9], and inpainting [18, 22] have seen significant progress due to their realistic generation results. [13] concentrated on creating full-body images by sampling from a trained texture-aware codebook, given human position and textual descriptions of clothing shapes and textures. Furthermore, in order to address the problem of pose-guided human prediction, [5] created a texture diffusion block that was conditioned by multi-scale texture patterns from the encoded source image. Adding to the tryon generation features, [4] introduced using the model pose, the garment sketch, and a textual description of the garment to condition the tryon generation process. Building on these methods and to improve the texture generation in person tryon, LaDI-VTON [20] utilised a textual inversion component, enabling mapping of garment visual features to the CLIP [23] token embedding space. This process generates a set of pseudo-word token embeddings, effectively conditioning the generation process. DCI-VTON [8] leverages a warping module to combine the warped clothes with clothes-agnostic person image and adds noise to guide the diffusion model's generation. Other methods on diffusion such as StableVITON [14] and CAT-DM [32] utilises a ControlNet based approach conditioned on straight garment for tryon.

3 Proposed Approach

Our model uses a two-stage pipeline. The first stage involves warping, with a graph-based warping module followed by a refinement module. It takes the source garment (I_g), a reference input (reference pose I_{pose} and agnostic image $I_{agnostic}$) as input. This stage computes dense flow f_o using graph correlation, producing a warped garment (I_{warp_g}) and a coarse try-on (I_{tryon_c}). The second stage generates the final try-on result using a diffusion model with an inpainting approach. Inputs include the person segmentation mask (I_{coarse_b}) from the coarse try-on, warped output (I_{warp_g}), human pose keypoints (I_{pose}), agnostic image ($I_{agnostic}$), and noise (I_z). The diffusion process is conditioned on the source cloth texture (I_g) and produces final try-on image(I_{tryon}). The diffusion process is conditioned with the attention based inversion between textual data

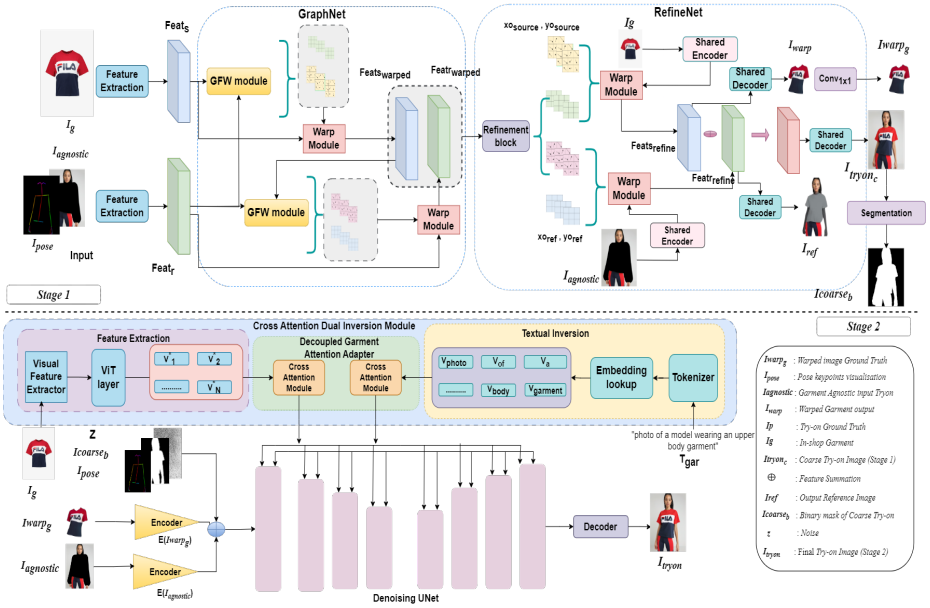

Fig. 1: Architecture Diagram of GraVITON. The top module utilizes GCNs for generating warped cloth and coarse tryon image. These outputs are processed to condition the Stable Diffusion model. The inversion model efficiently computes Cross-Modal attention to improve texture and structural consistency, generating the final tryon image.

(T_{gar}) and source cloth for texture (I_g). The calculated decoupled attention conditions the latent space to generate final tryon(I_{tryon}).

3.1 Graph based coarse tryon

The coarse tryon stage caters to the generation of warped garment I_{warp_g}) along with coarse tryon (I_{tryon_c}) that is further used in final tryon generation in stage 2. The input to the first stage, is source garment(I_g), reference pose (I_{pose}) and agnostic image ($I_{agnostic}$). The network employs a feature extraction module in form of convolution layers with N=3, N being the number of conv layers and a stride 2. The features extracted for both source($Feat_s$) and reference($Feat_r$) input are fed to the GraphNet module, that returns the warped source $Feat_{s_{warped}}$ and reference features $Feat_{r_{warped}}$ that are fed to the RefineNet for predicting final offsets $(x_{o_{source}}, y_{o_{source}})$ and $(x_{o_{ref}}, y_{o_{ref}})$ as shown in Figure 1.

GraphNet The overall working of GraphNet is similar to SDAFN [3], with the major difference is appearance flow estimation. The convolutional deformable flow warping stage in SDAFN is replaced by our novel Graph based Flow Warping (GFW) module as shown in Figure 2. The features extracted for both source ($Feat_s$) and reference ($Feat_r$) further act as an input to GFW module. The dense flow offsets $(x_{o_{source}}, y_{o_{source}})$ along with the computed attention maps are

utilised by the warping module to warp $Feat_s$ feature to compute source warped feature $Feat_{s_{warped}}$. Similarly, the source warped feature $Feat_{s_{warped}}$ and reference feature $Feat_r$ are fed to the GFW module to compute reference warped feature $Feat_{r_{warped}}$ from offsets $(x_{o_{ref}}, y_{o_{ref}})$.

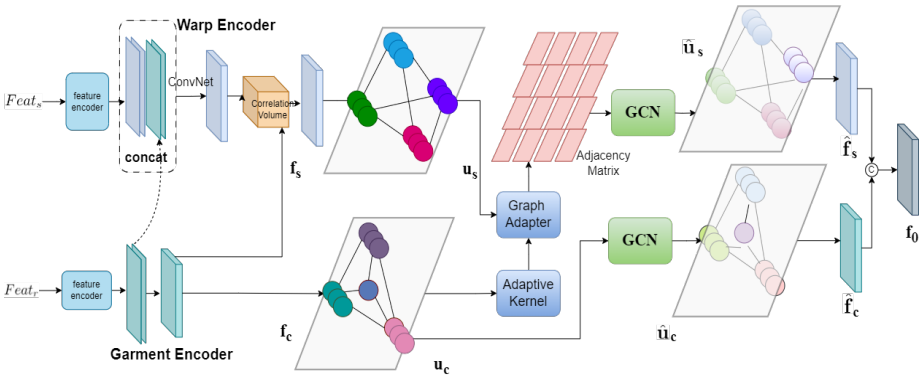

Fig. 2: Our Graph based Flow Warping (GFW) Module.

RefineNet The RefineNet module computes the refined offsets for predicting the warped garment I_{warp_g} and coarse tryon I_{tryon_c}.

The concatenated source and reference warped features $Feat_{s_{warped}}$ and $Feat_{r_{warped}}$ are fed as the input to refinement block to compute final offsets $x_{o_{source}}, y_{o_{source}}$ and $x_{o_{ref}}, y_{o_{ref}}$ which are the final warping directives for source garment (I_g) and reference input $(I_{agnostic})$ respectively. The refinement block is a simple four layer convolutional network based on [3]. The cloth I_g is sent to the shared encoder to compute garment features which are sent to the warp module along with the final source offsets $x_{o_{source}}, y_{o_{source}}$ to compute final garment warped feature $Feat_{s_{refine}}$. Similarly, the image agnostic $I_{agnostic}$ is sent to the shared encoder to compute agnostic features which are sent to the warp module along with the final reference offsets $x_{o_{ref}}, y_{o_{ref}}$ to compute warped reference feature $Feat_{r_{refine}}$. Both source and reference refined features are summed and sent to a shared decoder to compute the warped output tryon image(I_{tryon_c}). Similarly, the source refined feature $Feat_{s_{refine}}$ is fed to the shared decoder to compute generated warped garment image (I_{warp}), which is further refined by being passed through a 1x1 convolution layer to compute I_{warp_g}.

Graph based Flow Warping module(GFW) The graph network provides a highly connected space utilising the dense pixel context for appearance flow estimation. The source and reference features $Feat_s$, $Feat_r$ are sent to a shared feature encoder, whose corresponding output is then utilized to construct a 4D correlation volume capturing the statistical similarity between the two as shown in Figure 2. The resulting value is sent to four convolutions to capture source

feature $\mathbf{f}_s$. The reference feature ($Feat_r$) is fed to the garment encoder network to compute context feature $\mathbf{f}_c$ as shown in Figure 2. Both features are utilised to perform a holistic warp reasoning by computing offsets $\mathbf{f}_o = (x_o, y_o)$.

The graph based module in stage 1 consists of nodes(N) and edges(E) formulated in a directed graph as G=(N,E). The node embeddings are mapped to the graph space using a simple projection function, $\mathbf{u} = \mathcal{P}_{f\to u}(\mathbf{f})$, where $\mathbf{u}$ denotes the nodes in graph space, $\mathcal{P}$ is the projection function and $\mathbf{f}$ depicts the feature space. We define the nodes mapped into context (garment) feature $\mathbf{f}_c$ and warp feature $\mathbf{f}_s$ encoded as, $\mathbf{u}_c = (u_c^1, u_c^2, ..., u_c^n)$ and $\mathbf{u}_s = (u_s^1, u_s^2, ..., u_s^n)$, where $\mathbf{u}_c$ is context nodes for garment warping while $\mathbf{u}_s$ is the warp nodes computed from the normalized feature correlation between source and reference features $Feat_s$ and $Feat_r$ in the graph space.

The process of node creation for both the source and context entails the computation of the adjacency matrix, which measures the similarity between all nodes denoted as $\mathbf{u}_c$ and $\mathbf{u}_s$. To facilitate adaptive graph learning, we employ $\mathcal{L}()$ as a graph learner, comprising of a two-layer convolutional network with ReLU activation. The first layer focuses on channel-wise learning for $\mathbf{u}_s$, while the second layer introduces node-wise interaction learning, resulting in a refined node representation for the source denoted as $\hat{\mathbf{u}}_s^{(t)}$.

$$\breve{\mathbf{A}}_s = \mathcal{L}(\mathbf{u}_s; \Theta(\mathbf{u}_c)); \hat{\mathbf{u}}_s^{=} \mathcal{F}_{\texttt{AG}}(\mathbf{u}_s, \breve{\mathbf{A}}) \tag{1}$$

$$\hat{\mathbf{u}}_c = \mathcal{F}_{\texttt{Graph}}(\mathbf{u}_c, \mathbf{A}), \texttt{where } \mathbf{A} = \mathbf{u}_c^{\mathsf{T}} \mathbf{u}_c, \tag{2}$$

The final adjacency matrix for context and warp nodes is formulated in equation 1 giving the modified nodes for the source, with $\Theta()$ signifying a parameter learner and $\mathcal{F}_{\texttt{AG}}$ is adaptive graph learning function for warping. The context nodes are computed as in equation 2 where $\mathcal{F}_{\texttt{Graph}}$, is graph learner function in the Graph Adapter block defining the warping context.

The projection function $\mathcal{P}$ preserves the spatial details during the first(initial) conversion to the graph space, and utilising this, the modified nodes are projected back from graph to feature space using the projection function $\mathcal{P}$ as shown in equation 3 and equation 4, giving $\hat{\mathbf{f}}_c$ garment (context) and source warp feature $\hat{\mathbf{f}}_s$.

$$\hat{\mathbf{f}}_c = \mathbf{f}_c + h\mathcal{P}_{\mathrm{v}\to\mathrm{f}}(\hat{\mathbf{u}}_c), \tag{3}$$

where, h denotes a learnable parameter that is initialized as 0 and gradually performs a weighted sum. Similarly, the source warp feature $\hat{\mathbf{f}}_s$ is produced by

$$\hat{\mathbf{f}}_s = \mathbf{f}_s + l\mathcal{P}_{\mathrm{s}\to\mathrm{f}}(\hat{\mathbf{u}}_s). \tag{4}$$

where, l denotes a learnable parameter The resultant features are then concatenated to give the resulting offsets on the original grid from source image.

$$\mathbf{f_o} = (x_o^g, y_o^g) = (1 + F_{ch}(\hat{\mathbf{f}}_s)) * concat(\hat{\mathbf{f}}_c, \hat{\mathbf{f}}_s) \tag{5}$$

where, F_{ch} signifies the channel attention.

The overall loss for training stage 1 is defined below, where $L_{style}, L_{prec}, L_{L1}$ are style, perceptual and L1 losses.

$$\mathcal{L} = (\lambda_{L1}\mathcal{L}_{L1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style}) \tag{6}$$

3.2 Cross Modal Attention for Inversion

We utilise the coarse tryon output I_{tryon_c} from stage 1 to compute all the preprocessing inputs at stage 2 including person agnostic $I_{agnostic}$, binary person segmentation mask I_{coarse_b}, as well as pose keypoints I_{pose}. The preprocessed inputs go into the diffusion model for training.

Diffusion model : The model consists of a latent encoder **E** and latent decoder **D** block, from a pretrained VAE. A time conditioned U-net is used with a denoising parameter ϵ. The diffusion encoder takes in the warped garment I_{Warp_g} and person agnostic processed by a VAE encoder E giving the warped encoded garment $E(I_{Warp_g})$ and encoded person agnostic $E(I_{agnostic})$. The additional inputs: pose I_{pose}, mask I_{coarse_b} and noise z are resized to the latent size and concatenated.

The resulting inputs to the network are combined as:
$\beta = [Z; I_{coarse_b}; I_{pose}; E(I_{Warp_g}); E(I_{agnostic})]$ and used for latent learning. As virtual tryon aims to transfer the given warped garment to the person, it is treated as an inpainting task, inspired by [20]. The stable diffusion model is used as an in-painting approach where the latent space is conditioned with our DGAA adaptor. Our proposed framework focuses to inpaint the masked area, but instead of being guided by a TPS based warped garment, our diffusion model is guided by the warped garment computed from stage 1.

A CLIP encoder is employed for textual inversion which takes textual dataT_{gar} as an input. Similarly, input straight cloth I_g is fed to a pretrained variational encoder, and the features are fed to a ViT layer to compute texture feature for the same. The texture features from image are represented in CLIP token embedding space, similar to [20]. The token embeddings from the textual data acts as a textual prompt that guides the garment texture positioning. To enhance this, we introduce a Decoupled garment attention adaptor to condition the Denoising UNet giving realistic tryon results.

Decoupled Garment Attention Adaptor (DGAA) Although LaDI-VTON [20] enhances diffusion with inversion, it generates try-ons with erroneous texture details due to ineffective embedding of image features as they simply feed the concatenated features to the cross-attention layers. To address this, we propose the Decoupled Garment Attention Adaptor. Similar to [31], which improves conditioning in text-to-image generation, we utilise DGAA for conditioning the inpainting task of virtual tryon.

The textual features obtained from the CLIP embedding x_t are fed into the cross attention layer along with the query features z, given by latent. Hence, the cross-attention equation is given as,

Fig. 3: Qualitative results generated by proposed method in comparison with recent state-of-the-art approaches.

$$\mathbf{z}' = \text{Attention}(\alpha, \beta, \gamma) = \text{Softmax}(\frac{\alpha\beta^\top}{\sqrt{d}})\gamma, \tag{7}$$

where, $\alpha = zW_\alpha$, $\beta = x_iW_\beta$ and $\gamma = x_iW_\gamma$ are the query, key, and values matrices from the text features and W_β, W_γ. are the corresponding weight matrices. In DGAA, the cross attention layers for text features and garment features are separate. We add a new cross attention layer, for each cross attention layer in the original UNet model to insert garment features. Given the garment features g_i, the output of new cross attention $\mathbf{z}''$ is computed as follows:

$$\mathbf{z}'' = \text{Attention}(\alpha, \beta', \gamma') = \text{Softmax}(\frac{\alpha(\beta')^\top}{\sqrt{d}})\gamma', \tag{8}$$

where, $\alpha = zW_\alpha$, $\beta' = g_iW'_\beta$ and $\gamma' = g_iW'_\gamma$ are the query, key, and values matrices from the image features and W'_β, W'_γ. are the corresponding weight matrices.

We use the same query for image cross-attention as for text cross-attention. Consequently, we only need to add two paramemters W'_β and W'_γ for each cross-attention layer. In order to speed up the convergence, W'_β and W'_γ are initialized from W_β and W_γ.

Combining both the equations, 7 and 8 we get the final cross attention equation as below,

$$\begin{aligned} \mathbf{z}^{new} = \text{Softmax}(\frac{\alpha\beta^\top}{\sqrt{d}})\gamma + \text{Softmax}(\frac{\alpha(\beta')^\top}{\sqrt{d}})\gamma' \\ \text{where } \alpha = \mathbf{z}\mathbf{W}_\alpha, \beta = \boldsymbol{x}_t\mathbf{W}_\beta, \gamma = \boldsymbol{x}_t\mathbf{W}_\gamma, \\ \beta' = \boldsymbol{x}_i\mathbf{W}'_\beta, \gamma' = \boldsymbol{x}_i\mathbf{W}'_\gamma \end{aligned} \tag{9}$$

Here, W'_k and W'_v are the only trainable weights.

Loss: The diffusion model learns from the l1 loss function over noise as in [20].

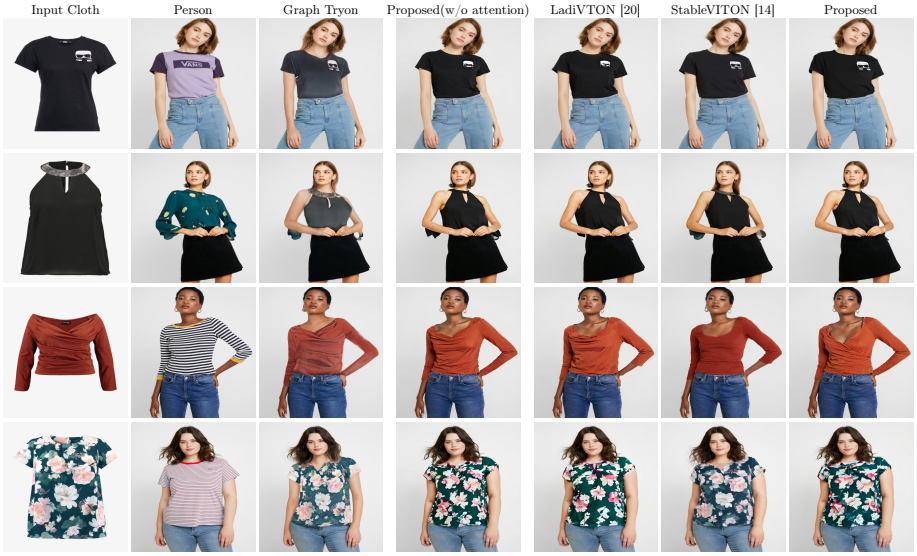

Fig. 4: Qualitative results showing the successive visual enhancement in results and analysis with LaDI-VTON and StableVITON.

3.3 Dataset

The experiments were conducted on VITON-HD and Dresscode datasets. VITON-HD is a high resolution dataset with resolution of 1024x768. The train set consists of 11,647 train pairs and 2,032 test pairs. DressCode is composed of 48,392/5,400 training/testing pairs of front-view full-body person and garment from different categories (i.e., upper, lower, dresses). The model is trained for both datasets in a paired setting on upper body garments and tested on both paired and unpaired setting. The same garment tryon is tested on the model as it is wearing in paired. While, a different garment tryon is tested on the model in an unpaired setting.

Fig. 5: Qualitative results of our proposed methodology on VITON-HD Dataset depicting pose, hair, sleeve length and texture variations.

Fig. 6: Qualitative results of our proposed methodology on Dresscode Dataset depicting pose, sleeve length, upper/dress and texture variations.

	SSIM	FID	KID
Graph based tryon	0.857	10.32	1.8
Flow based tryon	0.851	10.77	2.1
Diffusion with flow	0.873	9.21	1.2
Diffusion with graph	**0.881**	**8.37**	**0.81**

Table 1: Quantitative comparison between proposed method and incremental modules on VITON-HD dataset for paired setting

3.4 Implementation Details and Training

The model is trained in two stages successively. The graph based warping stage is trained first for 200 epochs, for a batch of 6 with a learning rate(LR) of 0.000035 on a V100 GPU. Weights for the loss functions are $\lambda_{L1} = 1, \lambda_{prec} = 1, \lambda_{style} = 100$. We used AdamW as training optimiser with $\beta 1 = 0.9$, $\beta 2 = 0.999$ and weight decay equal to 1e-2.

For training stage two, the inputs derived from stage 1 are utilised to construct the preprocessed inputs. This requires training our decoupled attention adaptor for 160 epochs with Adam optimizer, batch size 8 and $1e^{-5}$ LR. We employ SSIM, FID and KID metrics in both paired and unpaired settings for evaluation.

	SSIM	FID
Iterative flow	0.852	9.08
Single stage flow	0.874	8.91
Deformable flow	0.881	8.37

Table 2: Flow method comparison on VITON-HD dataset

3.5 Qualitative Results

To qualitatively assess our findings, we present sample images generated by our model alongside those by competing methods in Figure 3. While VITON-HD and HR-VITON have limitations in texture and warp accuracy, LaDI-VTON slightly improves texture details but looses colour consistency in garment. StableVITON doesnt preserve the garment shape accurately but improves colour and texture consistency. OOTDiffusion further improves texture details yet struggles to keep the garment person alignment intact as can be seen in row 2(right shoulder). Our

	SSIM	FID
Attention	0.881	8.37
W/O Attention	0.873	9.24

Table 3: Effect of attention in inversion module

approach produces highly realistic images, preserving the intricate textures and details of the original garments and garment warp using a decoupled attention-based inversion module and graph-based flow estimation.

In Figure 4, we compare both stages of our approach. Graph Tryon output from stage one shows initial garment warp, providing a baseline for refinement. In Stage 2, the diffusion model generates the warped garment on the agnostic image. This stage uses graph-based flow warping for preprocessing, generating rich textures and ensuring correct global warp. Through visual inspection, we discern improvements in texture preservation, micro-texture retention in green top(last row), spatial coherence in black dress(second row), and consistent boundary warp(third row) by our proposed approach. Figure 4 also compares our stage 1 and stage 2 results with the existing state-of-the-art methods LaDI-VTON and StableVITON. We observe that even without attention and inversion, our approach performs slightly better than LaDI-VTON. This is due to incorporation of our graph based warping stage, which predicts much better warps than TPS utilised in LaDI-VTON. The proposed approach as can be seen in last column, retains better texture and aligns garment optimally according to person's pose, thereby giving the best results.

Figure 5 depicts the garment tryon in an unpaired setting for VITON-HD dataset with texture variations, sleeve lengths, pose and hair. The generated images provide visual effectiveness of our method to handle self occlusion due to complex arm positions as can be seen in four images from the left. The proposed method also generates realistic garment textures retaining the fine details of text and symbols in the images. Figure 6 shows realistic tryon generation for Dresscode dataset. Our work generates realistic tryon for garments in unpaired setting. The results preserve texture, sleeve length and are agnostic to pose variations.

Graph	**Diffusion**	**Inversion**	**DGAA**	**SSIM**	**FID**
×	×	×	×	0.851	11.25
✓	×	×	×	0.857	10.32
✓	✓	×	×	0.868	9.78
✓	✓	✓	×	0.873	9.24
✓	✓	✓	✓	**0.881**	**8.37**

Table 4: Quantitative Ablation of our proposed modules on VITON-HD dataset

Table 5: Quantitative results on the VITON-HD dataset [6]. The best results are reported in **bold**.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
VITON-HD [6]	0.116	0.863	12.13	3.22
HR-VITON [16]	0.097	0.878	12.30	3.82
LaDI-VTON [20]	0.091	0.875	9.31	1.53
GP-VTON [33]	0.083	**0.892**	9.17	0.93
StableVITON [14]	0.084	0.862	9.13	1.20
OOTDiffusion [28]	0.071	0.878	8.81	0.82
Proposed	**0.070**	0.881	**8.37**	**0.81**

3.6 Quantitative results

We describe the robustness and correctness of our proposed approach by conducting extensive experiments and ablation on Dresscode and VITON-HD datasets. Table 1 demonstrates that the affect of introduction of graph for garment warping and coarse try-on prediction improves the accuracy of try-on module significantly when compared with the flow based traditional counterparts. It also describes the improvement in final try-on after utilising our diffusion model for target person generation. As we see, combination of Graph and Diffusion achieves the best result quantitatively.

Table 2 describes how various flow modules aid in warping input garment. The iterative flow which was motivated from RAFT [19] is unable to learn optimal warp, as the flow being learnt is an intermediate component of our network. While, RAFT [19] being a supervised framework introduced a flow consistency constraint which utilises ground truth flow that aids in learning of the iterative flow. We learn flow as an intermediate component in self-supervised manner. We also see that introduction of deformable flow [3] to our graph based flow estimation framework drastically improves learning of warped garment. This enhancement can be attributed to the fusion of features warped using multiple flows, resulting in the creation of a single optimized try-on. Consequently, while individual warped features may exhibit slight discrepancies, the fusion process aggregates the most favorable attributes from all features to generate an optimal try-on output.

The introduction of decoupled cross attention between text embedding and garment texture feature embedding improves the consistency of texture learnt in final tryon. This can be seen as improvement in FID and SSIM scores in table 3.

We analyzed the impact of each component on the performance of the model. As shown in Table 4, incorporating graph-based flow estimation led to a notable improvement in SSIM scores, indicating enhanced spatial coherence and perceptual quality in the generated images. Similarly, the integration of diffusion mechanisms in the generation process resulted in significantly lower FID scores, demonstrating improved fidelity and realism in the synthesized outputs.

Table 6: Quantitative results on the Dress Code dataset [21]. The best results are reported in **bold**.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
PSAD [21]	0.058	0.918	17.51	7.15
Paint-by-Example [29]	0.078	0.851	18.63	4.81
LaDI-VTON [20]	0.067	0.910	12.30	1.30
GP-VTON [33]	0.051	0.921	12.20	1.22
Proposed	**0.041**	**0.925**	**10.86**	**0.69**

The inclusion of attention mechanisms within the inversion module led to substantial gains in both SSIM and FID metrics, highlighting the importance of selective feature extraction and reconstruction in enhancing image quality and content preservation. Our comprehensive approach, combining all key components yielded the most impressive results.

As depicted in Table 5 and 6, we achieve highest SSIM and lowest FID scores among all prominent tryon methods on VITON-HD and Dresscode datasets, demonstrating the synergistic effects of our holistic technique.

4 Conclusion

Our paper introduces novel solutions to enhance virtual try-on, addressing critical challenges in garment warping and generation. By incorporating novel Graph-based Flow Warping module (GFW), we achieve accurate context reasoning, significantly reducing uncertainty in garment transfer. We introduce latent inversion for rich garment and text conditioning to a stable diffusion inpainting model. Our novel Decoupled Garment Cross-Attention Mechanism (DGAA) enriches latent space information of the diffusion model, leading to realistic try-on. Empirical validation on VITON-HD and DressCode datasets demonstrates substantial improvements in garment warping, texture preservation, and overall realism compared to existing methods.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
2. Bai, M., Luo, W., Kundu, K., Urtasun, R.: Exploiting semantic information and deep matching for optical flow. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 154–170. Springer (2016)
3. Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:250644446

4. Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. arXiv preprint arXiv:2304.02051 (2023)
5. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5968–5976 (2023)
6. Choi, S., Park, S., Lee, M.G., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14126–14135 (2021), https://api.semanticscholar.org/CorpusID:232427801
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. ArXiv **abs/2105.05233** (2021), https://api.semanticscholar.org/CorpusID:234357997
8. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia (2023)
9. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
10. Han, X., Huang, W., Hu, X., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10470–10479 (2019), https://api.semanticscholar.org/CorpusID:204959889
11. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7543–7552 (2017), https://api.semanticscholar.org/CorpusID:4532827
12. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3460–3469 (2022), https://api.semanticscholar.org/CorpusID:247939336
13. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation. ACM Transactions on Graphics (TOG) **41**(4), 1–11 (2022)
14. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8176–8185 (2024)
15. Lal Bhatnagar, B., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. arXiv e-prints pp. arXiv–1908 (2019)
16. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. arXiv preprint arXiv:2206.14180 (2022)
17. Li, Z., Wei, P., Yin, X., Ma, Z., Kot, A.C.: Virtual try-on with pose-garment keypoints guided inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22788–22797 (October 2023)
18. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
19. Luo, A., Yang, F., Luo, K., Li, X., Fan, H., Liu, S.: Learning optical flow with adaptive graph reasoning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 1890–1898 (2022)
20. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501 (2023)
21. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2230–2234 (2022), https://api.semanticscholar.org/CorpusID:248240016
22. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
24. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
25. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV). pp. 589–604 (2018)
26. Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M.C., Liang, X.: Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. In: Neural Information Processing Systems (2021), https://api.semanticscholar.org/CorpusID:244478414
27. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8121–8130 (2022)
28. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779 (2024)
29. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
30. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7850–7859 (2020)
31. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
32. Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., Liu, A.: Cat-dm: Controllable accelerated virtual try-on with diffusion model. arXiv preprint arXiv:2311.18405 (2023)
33. Zhenyu, X., Zaiyu, H., Xin, D., Fuwei, Z., Haoye, D., Xijin, Z., Feida, Z., Xiaodan, L.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)