TryOnDiffusion: A Tale of Two UNets

Luyang Zhu\textsuperscript{1,2}\textsuperscript{*}, Dawei Yang\textsuperscript{2}, Tyler Zhu\textsuperscript{2}, Fitsum Reda\textsuperscript{2}, William Chan\textsuperscript{2}, Chitwan Saharia\textsuperscript{2}, Mohammad Norouzi\textsuperscript{2}, Ira Kemelmacher-Shlizerman\textsuperscript{1,2}

\textsuperscript{1}University of Washington \textsuperscript{2}Google Research

\textsuperscript{*}Work done while author was an intern at Google.

Figure 1. TryOnDiffusion generates apparel try-on results with a significant body shape and pose modification, while preserving garment details at 1024×1024 resolution. Input images (target person and garment worn by another person) are shown in the corner of the results.

Abstract

Given two images depicting a person and a garment worn by another person, our goal is to generate a visualization of how the garment might look on the input person. A key challenge is to synthesize a photorealistic detail-preserving visualization of the garment, while warping the garment to accommodate a significant body pose and shape change across the subjects. Previous methods either focus on garment detail preservation without effective pose and shape variation, or allow try-on with the desired shape and pose but lack garment details. In this paper, we propose a diffusion-based architecture that unifies two UNets (referred to as Parallel-UNet), which allows us to preserve garment details and warp the garment for significant pose and body change in a single network. The key ideas behind Parallel-UNet include: 1) garment is warped implicitly via a cross attention mechanism, 2) garment warp and person blend happen as part of a unified process as opposed to a sequence of two separate tasks. Experimental results indicate that TryOnDiffusion achieves state-of-the-art performance.
both qualitatively and quantitatively.

1. Introduction

Virtual apparel try-on aims to visualize how a garment might look on a person based on an image of the person and an image of the garment. Virtual try-on has the potential to enhance the online shopping experience, but most try-on methods only perform well when body pose and shape variation is small. A key open problem is the non-rigid warping of a garment to fit a target body shape, while not introducing distortions in garment patterns and texture [5, 12, 41].

When pose or body shape vary significantly, garments need to warp in a way that wrinkles are created or flattened according to the new shape or occlusions. Related works [1, 5, 23] have been approaching the warping problem via first estimating pixel displacements, e.g., optical flow, followed by pixel warping, and postprocessing with perceptual loss when blending with the target person. Fundamentally, however, the sequence of finding displacements, warping, and blending often creates artifacts, since occluded parts and shape deformations are challenging to model accurately with pixel displacements. It is also challenging to remove those artifacts later in the blending stage even if it is done with a powerful generative model. As an alternative, TryOnGAN [24] showed how to warp without estimating displacements, via a conditional StyleGAN2 [21] network and optimizing in generated latent space. While the generated results were of impressive quality, outputs often lose details especially for highly patterned garments due to the low representation power of the latent space.

In this paper, we present TryOnDiffusion that can handle large occlusions, pose changes, and body shape changes, while preserving garment details at 1024×1024 resolution. TryOnDiffusion takes as input two images: a target person image, and an image of a garment worn by another person. It synthesizes as output the target person wearing the garment. The garment might be partially occluded by body parts or other garments, and requires significant deformation. Our method is trained on 4 Million image pairs. Each pair has the same person wearing the same garment but appears in different poses.

TryOnDiffusion is based on our novel architecture called Parallel-UNet consisting of two sub-UNets communicating through cross attentions [40]. Our two key design elements are implicit warping and combination of warp and blend (of target person and garment) in a single pass rather than in a sequential fashion. Implicit warping between the target person and the source garment is achieved via cross attention over their features at multiple pyramid levels which allows to establish long range correspondence. Long range correspondence performs well, especially under heavy occlusion and extreme pose differences. Furthermore, using the same network to perform warping and blending allows the two processes to exchange information at the feature level rather than at the color pixel level which proves to be essential in perceptual loss and style loss [19, 29]. We demonstrate the performance of these design choices in Sec. 4.

To generate high quality results at 1024×1024 resolution, we follow Imagen [35] and create cascaded diffusion models. Specifically, Parallel-UNet based diffusion is used for 128×128 and 256×256 resolutions. The 256×256 result is then fed to a super-resolution diffusion network to create the final 1024×1024 image.

In summary, the main contributions of our work are: 1) try-on synthesis at 1024×1024 resolution for a variety of complex body poses, allowing for diverse body shapes, while preserving garment details (including patterns, text, labels, etc.), 2) a novel architecture called Parallel-UNet, which can warp the garment implicitly with cross attention, in addition to warping and blending in a single network pass. We evaluated TryOnDiffusion quantitatively and qualitatively, compared to recent state-of-the-art methods, and performed an extensive user study. The user study was done by 15 non-experts, ranking more than 2K distinct random samples. The study showed that our results were chosen as the best 92.72% of the time compared to three recent state-of-the-art methods.

2. Related Work


Despite great progress, these methods still suffer from misalignment brought by explicit flow estimation and warping. TryOnGAN [24] tackles this issue by training a pose-conditioned StyleGAN2 [21] on unpaired fashion images and running optimization in the latent space to achieve try-on. By optimizing the latent space, however, it loses garment details that are less represented by the latent space. This becomes evident when garments have a pattern or details like pockets, or special sleeves.
We propose a novel architecture which performs implicit warping (without computing flow) and blending in a single network pass. Experiments show that our method can preserve details of the garment even under heavy occlusions and various body poses and shapes.

**Diffusion Models.** Diffusion models [15, 37, 39] have recently emerged as the most powerful family of generative models. Unlike GANs [4, 10], diffusion models have better training stability and mode coverage. They have achieved state-of-the-art results on various image generation tasks, such as super-resolution [36], colorization [34], novel-view synthesis [42] and text-to-image generation [28, 31, 33, 35]. Although being successful, state-of-the-art diffusion models utilize a traditional UNet architecture [15, 32] with channel-wise concatenation [34, 36] for image conditioning. The channel-wise concatenation works well for image-to-image translation problems where input and output pixels are perfectly aligned (e.g., super-resolution, inpainting and colorization). However, it is not directly applicable to our task as try-on involves highly non-linear transformations like garment warping. To solve this challenge, we propose Parallel-UNet architecture tailored to try-on, where the garment is warped implicitly via cross attentions.

**3. Method**

Fig. 2 provides an overview of our method for virtual try-on. Given an image \( I_p \) of person \( p \) and an image \( I_g \) of a different person in garment \( g \), our approach generates try-on result \( I_{tp} \) of person \( p \) wearing garment \( g \). Our method is trained on paired data where \( I_p \) and \( I_g \) are images of the same person wearing the same garment but in two different poses. During inference, \( I_p \) and \( I_g \) are set to images of two different people wearing different garments in different poses. We begin by describing our preprocessing steps, and a brief paragraph on diffusion models. Then we describe in subsections our contributions and design choices.
Preprocessing of inputs. We first predict human parsing map \((S_p, S_a)\) and 2D pose keypoints \((J_p, J_a)\) for both person and garment images using off-the-shelf methods [9,26]. For garment image, we further segment out the garment \(I_c\) using the parsing map. For person image, we generate clothing-agnostic RGB image \(I_a\) which removes the original clothing but retains the person identity. Note that clothing-agnostic RGB described in VITON-HD [5] leaks information of the original garment for challenging human poses and loose garments. We thus adopt a more aggressive way to remove the garment information. Specifically, we first mask out the whole bounding box area of the foreground person, and then copy-paste the head, hands and lower body part on top of it. We use \(S_p\) and \(J_p\) to extract the non-garment body parts. We also normalize pose keypoints to the range of \([0,1]\) before inputting them to our networks. Our try-on conditional inputs are denoted as \(c\text{tryon} = (I_a, J_p, I_c, J_g)\).

**Brief overview of diffusion models.** Diffusion models [15,37] are a class of generative models that learn the target distribution through an iterative denoising process. They consist of a Markovian forward process that gradually corrupts the data sample \(x\) into the Gaussian noise \(z_T\), and a learnable reverse process that converts \(z_T\) back to \(x\) iteratively. Diffusion models can be conditioned on various signals such as class labels, texts or images. A conditional diffusion model \(X_\theta\) can be trained with a weighted denoising score matching objective:

\[
\mathbb{E}_{x,c,\epsilon,t}[w_t \| X_\theta(\alpha_t x + \sigma_t \epsilon, c) - x\|^2] \tag{1}
\]

where \(x\) is the target data sample, \(c\) is the conditional input, \(\epsilon \sim \mathcal{N}(0, I)\) is the noise term, \(\alpha_t, \sigma_t, w_t\) are functions of the timestep \(t\) that affect sample quality. In practice, \(X_\theta\) is reparameterized as \(X_{\theta_\delta}\) to predict the noise that corrupts \(x\) into \(z_t := \alpha_t x + \sigma_t \epsilon\). At inference time, data samples can be generated from Gaussian noise \(z_T \sim \mathcal{N}(0, I)\) using samplers like DDPM [15] or DDIM [38].

3.1. Cascaded Diffusion Models for Try-On

Our cascaded diffusion models consist of one base diffusion model and two super-resolution (SR) diffusion models. The base diffusion model is parameterized as a \(128 \times 128\) Parallel-UNet (see Fig. 2 bottom). It predicts the \(128 \times 128\) try-on result \(I^{128}_{tr}\). Taking in the try-on conditional inputs \(c\text{tryon}\), since \(I_a\) and \(I_c\) can be noisy due to inaccurate human parsing and pose estimations, we apply noise conditioning augmentation [16] to them. Specifically, randomly Gaussian noise is added to \(I_a\) and \(I_c\) before any other processing. The levels of noise augmentation are also treated as conditional inputs following [16].

The \(128 \times 128\) \(\rightarrow 256 \times 256\) \(\rightarrow 256 \times 256\) SR diffusion model is parameterized as a \(256 \times 256\) Parallel-UNet. It generates the \(256 \times 256\) try-on result \(I^{256}_{tr}\) by conditioning on both the \(128 \times 128\) try-on result \(I^{128}_{tr}\) and the try-on conditional inputs \(c\text{tryon}\) at \(256 \times 256\) resolution. \(I^{256}_{tr}\) is directly downsampled from the ground-truth during training. At test time, it is set to the prediction from the base diffusion model. Noise conditioning augmentation is applied to all conditional input images at this stage, including \(I^{128}_{tr}, I_a\) and \(I_c\).

The \(256 \times 256\) \(\rightarrow 1024 \times 1024\) \(\rightarrow 1024 \times 1024\) SR diffusion model is parameterized as Efficient-UNet introduced by Imagen [35]. This stage is a pure super-resolution model, with no try-on conditioning. For training, random \(256 \times 256\) crops, from \(1024 \times 1024\), serve as the ground-truth, and the input is set to \(64 \times 64\) images downsampled from the crops. During inference, the model takes as input \(256 \times 256\) try-on result from previous Parallel-UNet model and synthesizes the final try-on result \(I^{256}_{tr}\) at \(1024 \times 1024\) resolution. To facilitate this setting, we make the network fully convolutional by removing all attention layers. Like the two previous models, noise conditioning augmentation is applied to the conditional input image.

3.2. Parallel-UNet

The \(128 \times 128\) Parallel-UNet can be represented as

\[
\epsilon_t = \epsilon_\theta(z_t, t, c\text{tryon}, t_{na}) \tag{2}
\]

where \(t\) is the diffusion timestep, \(z_t\) is the noisy image corrupted from the ground-truth at timestep \(t\), \(c\text{tryon}\) is the try-on conditional inputs, \(t_{na}\) is the set of noise augmentation levels for different conditional images, and \(\epsilon_t\) is predicted noise that can be used to recover the ground-truth from \(z_t\). The \(256 \times 256\) Parallel-UNet takes in the try-on result \(I^{128}_{tr}\) as input, in addition to the try-on conditional inputs \(c\text{tryon}\) at \(256 \times 256\) resolution. Next, we describe two key design elements of Parallel-UNet.

**Implicit warping.** The first question is: how can we implement implicit warping in the neural network? One natural solution is to use a traditional UNet [15,32] and concatenate the segmented garment \(I_c\) and the noisy image \(z_t\) along the channel dimension. However, channel-wise concatenation [34,36] can not handle complex transformations such as garment warping (see Sec. 4). This is because the computational primitives of the traditional UNet are spatial convolutions and spatial self attention, and these primitives have strong pixel-wise structural bias. To solve this challenge, we propose to achieve implicit warping using cross attention mechanism between our streams of information \((I_a\) and \(z_t)\). The cross attention is based on the scaled dot-product attention introduced by [40]:

\[
\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{3}
\]

where \(Q \in \mathbb{R}^{M \times d}, K \in \mathbb{R}^{N \times d}, V \in \mathbb{R}^{N \times d}\) are stacked vectors of query, key and value, \(M\) is the number of query
vectors, $N$ is the number of key and value vectors and $d$ is the dimension of the vector. In our case, the query and key-value pairs come from different inputs. Specifically, $Q$ is the flattened features of $z_t$ and $K, V$ are the flattened features of $I_c$. The attention map $\frac{QK^T}{\sqrt{d_k}}$ computed through dot-product tells us the similarity between the target person and the source garment, providing a learnable way to represent correspondence for the try-on task. We also make the cross attention multi-head, allowing the model to learn from different representation subspaces.

Combining warp and blend in a single pass. Instead of warping the garment to the target body and then blending with the target person as done by prior works, we combine the two operations into a single pass. As shown in Fig. 2, we achieve it via two UNets that handle the garment and the person respectively.

The person-UNet takes the clothing-agnostic RGB $I_o$ and the noisy image $z_t$ as input. Since $I_o$ and $z_t$ are pixel-wise aligned, we directly concatenate them along the channel dimension at the beginning of UNet processing.

The garment-UNet takes the segmented garment image $I_c$ as input. The garment features are fused to the target image via cross attentions defined above. To save model parameters, we early stop the garment-UNet after the $32 \times 32$ upsampling block, where the final cross attention module in person-UNet is done.

The person and garment poses are necessary for guiding the warp and blend process. They are first fed into the linear layers to compute pose embeddings separately. The pose embeddings are then fused to the person-UNet through the attention mechanism, which is implemented by concatenating pose embeddings to the key-value pairs of each self attention layer [35]. Besides, pose embeddings are reduced along the keypoints dimension using CLIP-style 1D attention pooling [27], and summed with the positional encoding of diffusion timestep $t$ and noise augmentation levels $t_{na}$. The resulting 1D embedding is used to modulate features for both UNets using FiLM [7] across all scales.

4. Experiments

Datasets. We collect a paired training dataset of 4 Million samples. Each sample consists of two images of the same person wearing the same garment in two different poses. For test, we collect 6K unpaired samples that are never seen during training. Each test sample includes two images of different people wearing different garments under different poses. Both training and test images are cropped and resized to $1024 \times 1024$ based on detected 2D human poses. Our dataset includes both men and women captured in different poses, with different body shapes, skin tones, and wearing a wide variety of garments with diverse texture patterns. In addition, we also provide results on the VITON-HD dataset [5].

Implementation details. All three models are trained with batch size 256 for 500K iterations using the Adam optimizer [22]. The learning rate linearly increases from 0 to $10^{-4}$ for the first 10K iterations and is kept constant afterwards. We follow classifier-free guidance [17] and train our models with conditioning dropout: conditional inputs are set to 0 for 10% of training time. All of our test results are generated with the following schedule: The base diffusion model is sampled for 256 steps using DDPM; The $128 \times 128 \rightarrow 256 \times 256$ SR diffusion model is sampled for 128 steps using DDPM; The final $256 \times 256 \rightarrow 1024 \times 1024$ SR diffusion model is sampled for 32 steps using DDIM. The guidance weight is set to 2 for all three stages. During training, levels of noise conditioning augmentation are sampled from uniform distribution $\mathcal{U}((0, 1))$. At inference time, they are set to constant values based on grid search, following [35].

Comparison with other methods. We compare our approach to three methods: TryOnGAN [24], SDAFN [2] and HR-VITON [23]. For fair comparison, we re-train all three methods on our 4 Million samples until convergence. Without re-training, the results of these methods are worse. Released checkpoints of SDAFN and HR-VITON also require layflat garment as input, which is not applicable to our setting. The resolutions of the related methods vary, and we present each method’s results in their native resolution: SDAFN’s at $256 \times 256$, TryOnGAN’s at $512 \times 512$ and HR-VITON at $1024 \times 1024$.

Quantitative comparison. Table 1 provides comparisons with two metrics. Since our test dataset is unpaired, we

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ours</th>
<th>VITON-HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>FID ↓ KID ↓</td>
<td>30.20% 18.586</td>
<td></td>
</tr>
<tr>
<td>TryOnGAN [24]</td>
<td>24.57% 16.024</td>
<td></td>
</tr>
<tr>
<td>HR-VITON [23]</td>
<td>18.705 9.200</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Random</th>
<th>Challenging</th>
</tr>
</thead>
<tbody>
<tr>
<td>TryOnGAN [24]</td>
<td>1.75%  0.45%</td>
<td></td>
</tr>
<tr>
<td>SDAFN [2]</td>
<td>2.42%  2.20%</td>
<td></td>
</tr>
<tr>
<td>HR-VITON [23]</td>
<td>2.92%  1.30%</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>92.72% 95.80%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Quantitative comparison to 3 baselines. We compute FID and KID on our 6K test set and VITON-HD’s unpaired test set. The KID is scaled by 1000 following [20].
Figure 3. Comparison with TryOnGAN [24], SDAFN [2] and HR-VITON [23]. First column shows the input (person, garment) pairs. TryOnDiffusion warps well garment details including text and geometric patterns even under extreme body pose and shape changes.

compute Frechet Inception Distance (FID) [14] and Kernel Inception Distance (KID) [3] as evaluation metrics. We computed those metrics on both test datasets (our 6K, and VITON-HD) and observe a significantly better performance with our method.

User study. We ran two user studies to objectively evaluate our methods compared to others at scale. The results are reported in Table 2. In first study (named “random”), we randomly selected 2804 input pairs out of the 6K test set, ran all four methods on those pairs, and presented to raters. 15 non-expert raters (on crowdsource platform) have been asked to select the best result out of four or choose “hard to tell” option. Our method was selected as best for 92.72% of the inputs. In a second study (named “challenging”), we performed the same setup but chose 2K input pairs (out of 6K) with more challenging poses. The raters selected our method as best for 95.8% of the inputs.

Qualitative comparison. In Figures 3 and 4, we provide visual comparisons to all baselines on two test datasets (our 6K, and VITON-HD). Note that many of the chosen input pairs have quite different body poses, shapes and complex garment materials—all limitations of most previous methods—thus we don’t expect them to perform well but present here to show the strength of our method. Specif-
Figure 4. Comparison with state-of-the-art methods on VITON-HD dataset [5]. All methods were trained on the same 4M dataset and tested on VITON-HD.

Figure 5. Qualitative results for ablation studies. Left: cross attention versus concatenation for implicit warping. Right: One network versus two networks for warping and blending. Zoom in to see differences highlighted by green boxes.

Typically, we observe that TryOnGAN struggles to retain the texture pattern of the garments while SDAFN and HR-VITON introduce warping artifacts in the try-on results. In contrast, our approach preserves fine details of the source garment and seamlessly blends the garment with the person even if the poses are hard or materials are complex (Fig. 3, row 4). Note also how TryOnDiffusion generates realistic garment wrinkles corresponding to the new body poses (Fig. 3, row 1). We show easier poses in the supplementary (in addition to more results) to provide a fair comparison to other methods.

**Ablation 1: Cross attention vs concatenation for implicit warping.** The implementation of cross attention is detailed in Sec. 3.2. For concatenation, we discard the garment-UNet, directly concatenate the segmented garment $I_c$ to the noisy image $z_t$, and drop cross attention modules in the person-UNet. We apply these changes to each Parallel-UNet, and keep the final SR diffusion model same. Fig. 5 shows that cross attention is better at preserving garment details under significant body pose and shape changes.

**Ablation 2: Combining warp and blend vs sequencing two tasks.** Our method combines both steps in one network pass as described in Sec. 3.2. For the ablated version, we train two base diffusion models while SR diffusion models are intact. The first base diffusion model handles the warping task. It takes as input the segmented garment $I_c$, the person pose $J_p$ and the garment pose $J_g$, and predicts the warped garment $I_{wc}$. The second base diffusion model performs the blending task, whose inputs are the warped garment $I_{wc}$, clothing-agnostic RGB $I_a$, person pose $J_p$ and garment pose $J_g$. The output is the try-on result $I_{tr}^{128}$ at 128×128 resolution. The conditioning for $(I_c, I_a, J_p, J_g)$ is kept unchanged. $I_{wc}$ in the second base diffusion model is processed by a garment-UNet, which is the same as $I_c$. Fig. 5 visualizes the results of both methods. We can see that sequencing warp and blend causes artifacts near the garment boundary, while a single network can blend the target person and the source garment nicely.
Figure 6. Failures happen due to erroneous garment segmentation (left) or garment leaks into the Clothing-agnostic RGB image (right).

Figure 7. TryOnDiffusion on eight target people (columns) dressed by five garments (rows). Zoom in to see details.

**Limitations.** First, our method exhibits garment leaking artifacts in case of errors in segmentation maps and pose estimations during preprocessing. Fortunately, those [9, 26] became quite accurate in recent years and this does not happen often. Second, representing identity via clothing-agnostic RGB is not ideal, since sometimes it may preserve only part of the identity, e.g., tatoos won’t be visible in this representation, or specific muscle structure. Third, our train and test datasets have mostly clean uniform background so it is unknown how the method performs with more complex backgrounds. Finally, this work focused on upper body clothing and we have not experimented with full body try-on, which is left for future work. Fig. 6 demonstrates failure cases.

Finally, Fig. 7 shows TryOnDiffusion results on variety of people and garments. Please refer to supplementary material for more results.

5. Summary and Future Work

We presented a method that allows to synthesize try-on given an image of a person and an image of a garment. Our results are overwhelmingly better than state-of-the-art, both in the quality of the warp to new body shapes and poses, and in the preservation of the garment. Our novel architecture Parallel-UNet, where two UNets are trained in parallel and one UNet sends information to the other via cross attentions, turned out to create state-of-the-art results. In addition to the exciting progress for the specific application of virtual try-on, we believe this architecture is going to be impactful for the general case of image editing, which we are excited to explore in the future. Finally, we believe that the architecture could also be extended to videos, which we also plan to pursue in the future.
References


