TryOnDiffusion: A Tale of Two UNets

Luyang Zhu[1,2*] Dawei Yang[2] Tyler Zhu[2] Fitsum Reda[2] William Chan[2]
Chitwan Saharia[2] Mohammad Norouzi[2] Ira Kemelmacher-Shlizerman[1,2]
[1]University of Washington [2]Google Research

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

[1]Work done while author was an intern at Google.

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

Image-Based Virtual Try-On. Given a pair of images (target person, source garment), image-based virtual try-on methods generate the look of the target person wearing the source garment. Most of these methods [2, 5, 6, 8, 12, 13, 18, 23, 25, 30, 41, 43–46] decompose the try-on task into two stages: a warping model and a blending model. The seminal work VITON [12] proposes a coarse-to-fine pipeline guided by the thin-plate-spline (TPS) warping of source garments. ClothFlow [11] directly estimates flow fields with a neural network instead of TPS for better garment warping. VITON-HD [5] introduces alignment-aware generator to increase the try-on resolution from 256×192 to 1024×768. HR-VITON [23] further improves VITON-HD by predicting segmentation and flow simultaneously. SDAFN [2] predicts multiple flow fields for both the garment and the person, and combines warped features through deformable attention [47] to improve quality.

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.

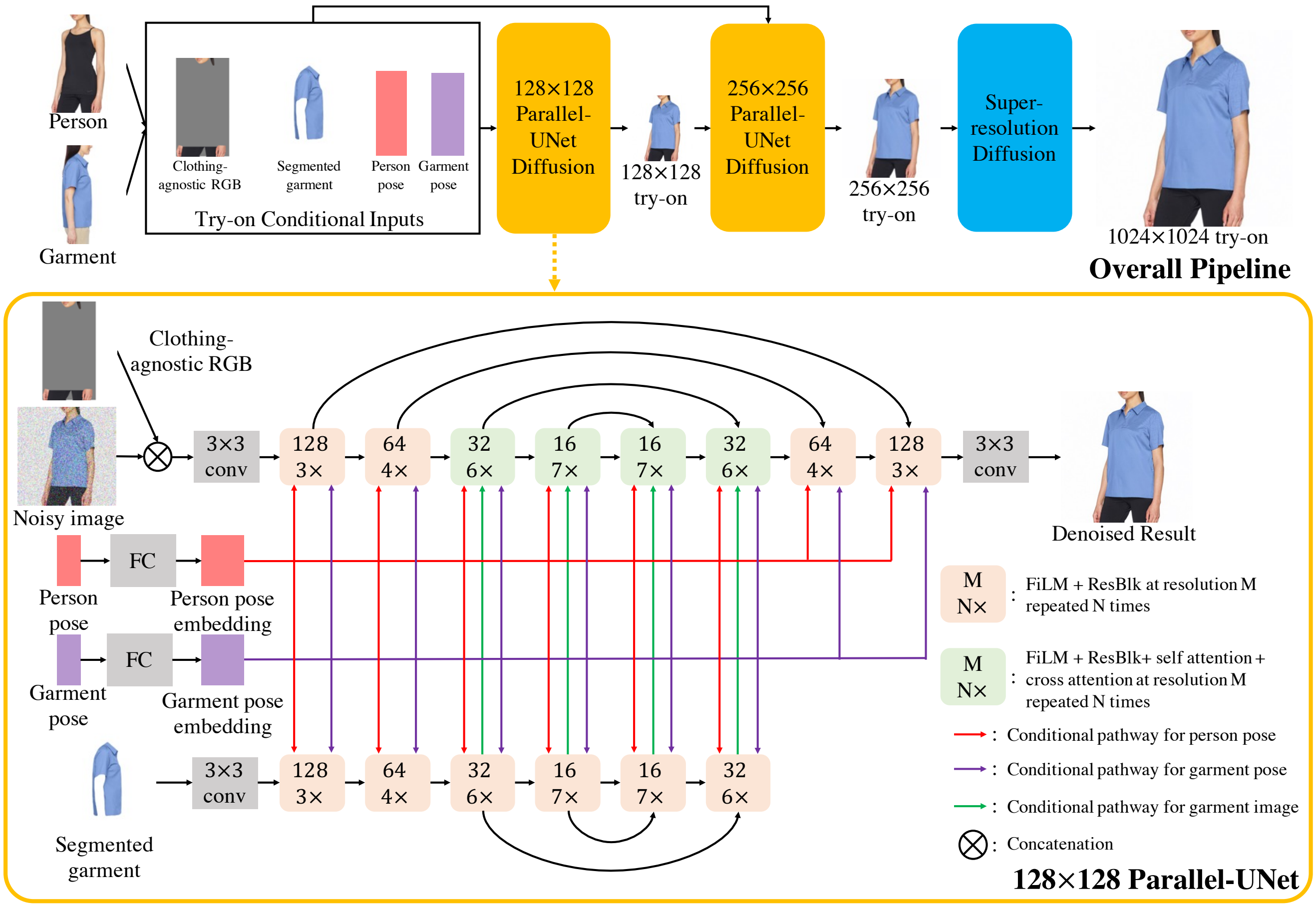

Figure 2. Overall pipeline (top): During preprocessing step, the target person is segmented out of the person image creating "clothing agnostic RGB" image, the target garment is segmented out of the garment image, and pose is computed for both person and garment images. These inputs are taken into 128×128 Parallel-UNet (key contribution) to create the 128 × 128 try-on image which is further sent as input to the 256×256 Parallel-UNet together with the try-on conditional inputs. Output from 256×256 Parallel-UNet is sent to standard super resolution diffusion to create the 1024×1024 image. The architecture of 128×128 Parallel-UNet is visualized at the bottom, see text for details. The 256×256 Parallel-UNet is similar to the 128 one, and provided in supplementary for completeness.

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_{tr} 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_g) and 2D pose keypoints (J_p, J_g) 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 $\mathbf{c}_{\text{tryon}} = (I_a, J_p, I_c, J_g)$.

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 $\mathbf{x}$ into the Gaussian noise $\mathbf{z}_T$, and a learnable reverse process that converts $\mathbf{z}_T$ back to $\mathbf{x}$ iteratively. Diffusion models can be conditioned on various signals such as class labels, texts or images. A conditional diffusion model $\hat{\mathbf{x}}_\theta$ can be trained with a weighted denoising score matching objective:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2] \tag{1}$$

where $\mathbf{x}$ is the target data sample, $\mathbf{c}$ is the conditional input, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise term. α_t, σ_t, w_t are functions of the timestep t that affect sample quality. In practice, $\hat{\mathbf{x}}_\theta$ is reparameterized as $\hat{\boldsymbol{\epsilon}}_\theta$ to predict the noise that corrupts $\mathbf{x}$ into $\mathbf{z}_t := \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$. At inference time, data samples can be generated from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 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 128×128 Parallel-UNet (see Fig. 2 bottom). It predicts the 128×128 try-on result I_{tr}^{128}, taking in the try-on conditional inputs $\mathbf{c}_{\text{tryon}}$. 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, random 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 $128\times128 \rightarrow 256\times256$ SR diffusion model is parameterized as a 256×256 Parallel-UNet. It generates the 256×256 try-on result I_{tr}^{256} by conditioning on both the 128×128 try-on result I_{tr}^{128} and the try-on conditional inputs $\mathbf{c}_{\text{tryon}}$ at 256×256 resolution. I_{tr}^{128} 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_{tr}^{128}, I_a and I_c.

The $256\times256 \rightarrow 1024\times1024$ 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 256×256 crops, from 1024×1024, serve as the ground-truth, and the input is set to 64×64 images downsampled from the crops. During inference, the model takes as input 256×256 try-on result from previous Parallel-UNet model and synthesizes the final try-on result I_{tr} at 1024×1024 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 128×128 Parallel-UNet can be represented as

$$\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}_{\text{tryon}}, \mathbf{t}_{\text{na}}) \tag{2}$$

where t is the diffusion timestep, $\mathbf{z}_t$ is the noisy image corrupted from the ground-truth at timestep t, $\mathbf{c}_{\text{tryon}}$ is the try-on conditional inputs, $\mathbf{t}_{\text{na}}$ is the set of noise augmentation levels for different conditional images, and $\boldsymbol{\epsilon}_t$ is predicted noise that can be used to recover the ground-truth from $\mathbf{z}_t$. The 256×256 Parallel-UNet takes in the try-on result I_{tr}^{128} as input, in addition to the try-on conditional inputs $\mathbf{c}_{\text{tryon}}$ at 256×256 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 $\mathbf{z}_t$ 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_c and $\mathbf{z}_t$). 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

Test datasets	Ours		VITON-HD	
Methods	FID ↓	KID ↓	FID ↓	KID ↓
TryOnGAN [24]	24.577	16.024	30.202	18.586
SDAFN [2]	18.466	10.877	33.511	20.929
HR-VITON [23]	18.705	9.200	30.458	17.257
Ours	**13.447**	**6.964**	**23.352**	**10.838**

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].

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 $\mathbf{z}_t$ 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_a and the noisy image $\mathbf{z}_t$ as input. Since I_a and $\mathbf{z}_t$ 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 32×32 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 $\mathbf{t}_{\text{na}}$. 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 1024×1024 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].

Methods	Random	Challenging
TryOnGAN [24]	1.75%	0.45%
SDAFN [2]	2.42%	2.20%
HR-VITON [23]	2.92%	1.30%
Ours	**92.72%**	**95.80%**
Hard to tell	0.18%	0.25%

Table 2. Two user studies. "Random": 2804 random input pairs (out of 6K) were rated by 15 non-experts asked to select the best result or choose "hard to tell". "Challenging": 2K pairs with challenging body poses were selected out of 6K and rated in same fashion. Our method significantly outperforms others in both studies.

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 $128\times128\rightarrow256\times256$ SR diffusion model is sampled for 128 steps using DDPM; The final $256\times256\rightarrow1024\times1024$ 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 256×256, TryOnGAN's at 512×512 and HR-VITON at 1024×1024.

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

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.

ically, 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 $\mathbf{z}_t$, 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.*, tatooes 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

[1] Walmart Virtual Try-On. https://www.walmart.com/cp/virtual-try-on/4879497. 2

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 2, 5, 6

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 2, 4, 5, 7

[6] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022. 2

[7] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018. 5

[8] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 2

[9] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 4, 8

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[11] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019. 2

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[13] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3470–3479, June 2022. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 4

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 4

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[18] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision*, pages 619–635. Springer, 2020. 2

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 5

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2, 5, 6

[24] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 2, 5, 6

[25] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 2

[26] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017. 4, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[29] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *The European Conference on Computer Vision (ECCV)*, 2022. 2

[30] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3, 4

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3

[34] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3, 4

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 2, 3, 4, 5

[36] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 4

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3, 4

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[39] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[41] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 2

[42] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[43] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, June 2022. 2

[44] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 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*, pages 7850–7859, 2020. 2

[45] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10511–10520, 2019. 2

[46] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021. 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2