UniHuman: A Unified Model For Editing Human Images in the Wild

Nannan Li[1]* Qing Liu[2] Krishna Kumar Singh[2] Yilin Wang[2]
Jianming Zhang[2] Bryan A. Plummer[1] Zhe Lin[2]
[1]Boston University [2]Adobe
nnli@bu.edu {qingl,krishsin,yilwang,jianmzha}@adobe.com
bplum@bu.edu zlin@adobe.com

Figure 1. The results of UniHuman on diverse real-world images. UniHuman learns informative representations by leveraging multiple data sources and connections between related tasks, achieving high-quality results across various human image editing objectives.

Abstract

Human image editing includes tasks like changing a person's pose, their clothing, or editing the image according to a text prompt. However, prior work often tackles these tasks separately, overlooking the benefit of mutual reinforcement from learning them jointly. In this paper, we propose UniHuman, a unified model that addresses multiple facets of human image editing in real-world settings. To enhance the model's generation quality and generalization capacity, we leverage guidance from human visual encoders and introduce a lightweight pose-warping module that can exploit different pose representations, accommodating unseen textures and patterns. Furthermore, to bridge the disparity between existing human editing benchmarks with real-world data, we curated 400K high-quality human image-text pairs for training and collected 2K human images for out-of-domain testing, both encompassing diverse clothing styles, backgrounds, and age groups. Experiments on both in-domain and out-of-domain test sets demonstrate that UniHuman outperforms task-specific models by a significant margin. In user studies, UniHuman is preferred by the users in an average of 77% of cases. Our project is available at this link.

1. Introduction

In the realm of computer graphics and computer vision, the synthesis and manipulation of human images have evolved into a captivating and transformative field. This field holds invaluable applications covering a range of domains: reposing strives to generate a new pose of a person given a target pose [2, 43, 45, 47], virtual try-on aims to seamlessly fit a new garment onto a person [23, 26, 48], and text-to-image editing manipulate a person's clothing styles based on text prompts [5, 11, 12, 40]. However, most approaches address these tasks in isolation, neglecting the benefits of learning them jointly to mutually reinforce one another via the utilization of auxiliary information provided by related tasks [9, 16, 42]. In addition, few studies have explored effective ways to adapt to unseen human-in-the-wild cases.

In response to these challenges, our goal is to unify multiple human image editing tasks in a single model, boosting performance in all settings. Thus, we propose UniHuman, a unified model that exploits the synergies among reposing, virtual try-on, and text manipulation for in-the-wild human image editing. Fig. 1 shows UniHuman's high-quality generations can both preserve the clothing identity and recover texture details. While prior work [7] made initial attempts to bridge the visual and text domains, connections within visual tasks remain largely under-explored. In contrast, our

*Work during Nannan Li's 2023 summer internship at Adobe Research.

model takes a step further by exploiting the relationship between reposing and virtual try-on. Specifically, reposing requires modifying the pose of all body parts and clothing items, while virtual try-on only adapts the pose of the target garment. Importantly, both tasks need to keep the visible texture consistent, either from the source image or the target garment, consistent after the pose or garment change. This suggests learning the two tasks could benefit each other when consolidated in a unified model. Moreover, when adding text prompts [26], the semantic information learned from large language models [32] could help our model synthesize more realistic textures based on the visible texture.

Recognizing that all three tasks should maintain the consistency of visible content, we introduce a pose-warping module. Instead of being a task-specific module trained from a single format of pose representation [1, 8, 13, 20, 23, 39, 41], our pose-warping module can explicitly leverage both dense and sparse pose correspondences to obtain visible pixels on all three tasks, equipping it with the capacity to handle previously unseen textures and patterns. In parallel, to maximize the utilization of human-specific visual information, we leverage human visual encoders at both the part and global levels to infuse texture, style, and pose guidance in the generation process. The visual representations are then taken as reference in the unified model to reconstruct the person after the desired pose/garment change. Our experiments show that the introduced pose-warping module can enhance the model's generalization capacity as well as generation quality at inference.

In order to adapt our model to real-world scenarios, it is essential to have ample data for learning informative representations across diverse domains. However, existing datasets are constrained by the scale and diversity [8, 18, 24, 25], particularly in terms of pose, background, and age groups. Relying solely on such data for training may introduce biases in image generation and hinder the model's ability to generalize to real-world samples. Additionally, existing in-the-wild human image/video datasets [17, 24] are marred by motion blur or by a lack of effective way to handle occlusions. Instead, we developed an automated data-curation pipeline to collect 400K high-quality human image-text pairs from LAION-400M [33], assembling images with at least 512x512 resolution as well as little to no human-object occlusion. This cost-effective new dataset contains single-human images with an extensive range of poses, clothing styles, backgrounds, *etc*. By jointly training on this dataset with existing datasets, we effectively enhance our model's generalization capacity on human-in-the-wild cases. Our contributions are:

- We propose UniHuman, a unified model that can achieve multiple in-the-wild human image editing tasks through human visual encoders. The incorporation of the pose-warping module further enhances the model's generalization capability at inference.
- We curated 400K high-quality human image-text pairs for training and collected 2K human image pairs for out-of-domain testing. Our data expands existing data and includes more diverse poses, backgrounds, and age groups.
- Extensive experiments on both in-domain and out-of-domain test sets show that UniHuman outperforms task-specific models quantitatively and qualitatively, and is preferred by the users in 77% cases on average.

2. Related Work

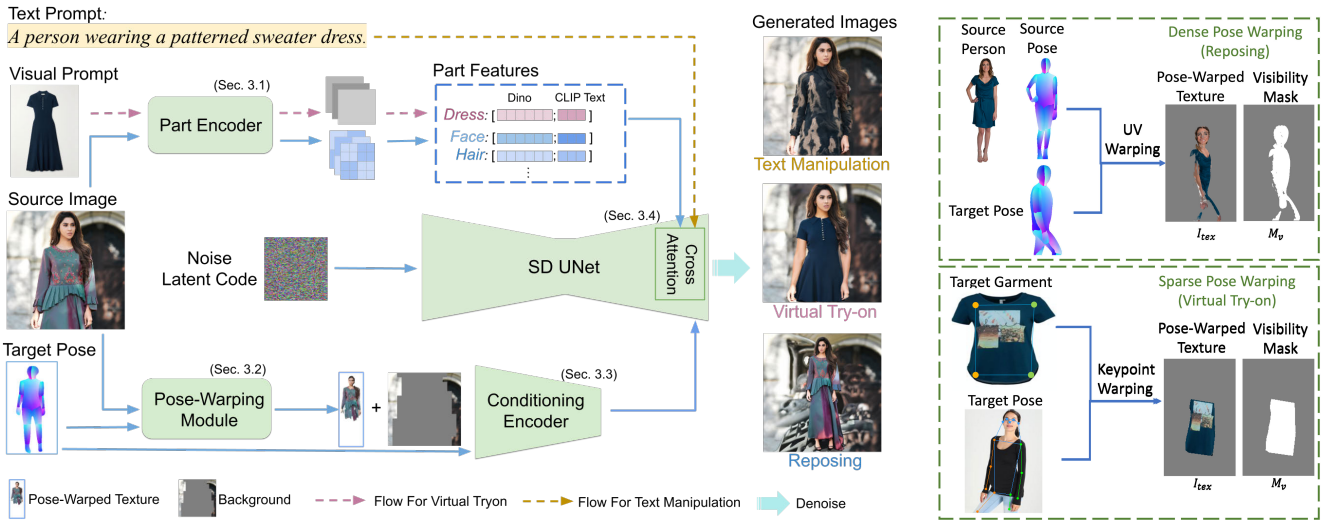
Human image editing empowers us to reshape, recolor, and infuse texture patterns into visual content. This paper tackles three distinct yet interrelated tasks: reposing, virtual try-on, and text manipulation. In contrast to methods that handle these tasks independently [1, 2, 5, 7, 10–12, 19–21, 23, 26, 34, 36, 38–40, 45, 48], our unified model, UniHuman, leverages synergies and mutual benefits from related tasks through joint learning. Additionally, we augment existing datasets with additional diverse and high-quality data to enable our model's adaptability to unseen human-in-the-wild cases, a less-explored setting in prior research.

Previous attempts at unifying the three tasks [7] used frozen CLIP embeddings [29] to bridge visual and text domains. However, the connections between different *visual* tasks in this setting still remained largely overlooked. In contrast, our model unifies all three human image editing tasks by introducing a pose-warping module. This module provides texture details derived from pose aligning, equipping it with the ability to handle unseen clothing patterns. While other methods [1, 8, 13, 20, 23, 39, 41] often rely on a single format of pose representation to train task-specific pose warping modules, UniHuman flexibly utilizes both dense and sparse pose representations for training-free texture warping on all three tasks. This approach mitigates the overfitting to specific datasets, thus ensuring accurate visible pixel information resilient to domain shifts. More related literature discussions can be found in Supp.

3. Human Image Editing in a Unified Model

Given a source image I_{src}, a target pose P_{tgt}, an optional visual prompt G_t and/or an optional text prompt y, our goal is to reconstruct the person of I_{src} at the target pose P_{tgt}, and simultaneously transfer the texture from G_t and/or create a new texture based on y. Note that we unify three human image editing tasks in one single model as follows: In the absence of a visual/text prompt, our focus narrows down to human reposing; when the visual prompt specifies a target garment, the task transforms into virtual try-on; conversely, when guided by a text description, the model edits the image following the principle of text manipulation.

Our UniHuman model is implemented based on Sta-



(a) Inference pipeline (Sec. 3). (b) Pose-Warping module. (Sec. 3.2)

Figure 2. An overview of our model. **(a)** Our inference pipeline. Starting from a noise latent code, our model edits the source person given the source image, the target pose, the visual prompt (optional), and the text prompt (optional). Blue arrow is the reposing flow, which is also the base flow for all tasks. Pink dashed arrow indicates the optional virtual try-on flow that takes a clothing image as its input. In *try-on* task, the clothing image should replace the source image as the input to the pose-warping module. Brown dashed arrow is the optional text manipulation flow, which accepts a text description as its prompt. **(b)** The introduced pose-warping module. It maps the original RGB pixels of the source texture to the target pose based on pose correspondences. Best view in color.

ble Diffusion (SD) [32]. The part encoder learns texture styles from segmented human parts, supplying them to SD cross-attention (Sec. 3.1). Simultaneously, the pose-warping module generates detailed, target pose-aligned visible texture (Sec. 3.2). These outputs, along with the target pose and partial background, channel into the SD UNet via a conditioning encoder (Sec. 3.3). For virtual try-on (Sec. 3.4), the optional target garment is injected into the part encoder to be combined with other human parts. In cases of text manipulation, the SD UNet learns semantic information from an optional text prompt (Sec. 3.4). After N-timestep denoising and VAE decoding, the model produces a clean edited image. Objective functions used in training are detailed in Sec. 3.5.

3.1. Part Encoder

To acquire texture information from the source person, we utilize a part encoder to obtain segmented human part features, which are then fed into the SD UNet decoder. Unlike the approach in Cheong et al. [7], where human parts are segmented at the pixel level and encoded separately, we segment parts at the feature level, *i.e.*, take parts from the feature map of the entire source person. We find this segmented feature map can preserve more contextual information than image segments, *e.g.*, the length of the clothing and interactions between the upper & lower clothing. An off-the-shelf human parsing model [22] is used to extract *face, hair, headwear, upper clothing, coat, lower clothing, shoes, accessories*, and *person* from the source person's DINOv2 [28] feature map. These visual features are then concatenated with the corresponding CLIP text embeddings. Let d_ω be the part encoder that includes DINOv2 and CLIP. The obtained part features $B = d_\omega(I_{src})$ will provide source texture and style information in the SD UNet in Sec. 3.4.

3.2. Pose-Warping Module

To ensure texture consistency after pose/garment change and improve our model's ability to generalize to unseen textures, we introduce a pose-warping module. This module produces the pose-warped texture I_{tex} and its binary mask M_v, which will be sent to the conditioning encoder in Sec. 3.3 and to the SD UNet cross-attention in Sec. 3.4. While earlier approaches [1, 13, 20, 23, 23, 39, 41] train task-specific pose warping modules, our model obtains the pose-warped texture through explicit correspondence mapping as shown in Fig. 2b. This process only needs an off-the-shelf pose detector to provide sparse or dense pose prediction for texture warping. Consequently, our method is inherently more resilient to domain shifts across different tasks, achieving enhanced generalization capacity to handle unseen patterns and styles.

For tasks involving human pose change, the pose-warped texture $I_{tex\text{-}rp}$ pertains to pixels that remain visible after reposing, as shown in Fig. 2b top panel. We use UV map correspondence to resample source RGB pixels such that they are aligned with the target pose. This alignment is

critical in enabling direct reconstruction of intricate texture patterns. However, in cases where only the target garment requires repositioning (as in virtual try-on), since no 3D or contextual information is available from the target garment image, warping the texture through UV coordinates becomes unfeasible. In such scenarios, as illustrated in Fig. 2b bottom panel, we pivot to the use of sparse keypoints to apply a perspective warping from the canonical view of the target garment to the human torso. This warping repositions the clothing texture to the desired pose, providing the pose-warped texture $I_{tex\text{-}vt}$ for virtual try-on. For text manipulation, the pose-warped texture $I_{tex\text{-}tm}$ exhibits adaptability, catering to user-specific requirements. For example, it can be set to zero to facilitate the generation of clothing textures from scratch based on the text input. Our experiments show that the introduced pose-warped texture strengthens the generalization capacity of our approach.

3.3. Conditioning Encoder

The conditioning encoder takes the target pose P_{tgt}, pose-warped texture I_{tex} and partial background I_{bg} as input, which provides essential posture guidance and visible texture reference for all tasks. The partial background image I_{bg} is extracted by masking out the bounding boxes of the source and target pose region. Following [44], the encoded features in g_ϕ are concatenated with the intermediate features in SD UNet decoder as follows

$$\hat{h}^i = W_h^i[h^i; g_\phi^i([I_{tex}; P_{tgt}; I_{bg}])], \tag{1}$$

where h^i is the i^{th} intermediate feature map of the SD UNet decoder, g_ϕ^i is the i^{th} intermediate layer of g_ϕ, W_h^i are learnable weights, and $[\cdot;\cdot]$ indicates concatenation. To obtain $\hat{h}^i$, the output feature maps of g_ϕ at varying resolutions are injected into every block of the SD UNet decoder. We define $E = g_\phi([I_{tex}; \emptyset; \emptyset])$ as the encoded pose-warped texture by itself in the last layer of g_ϕ, which will be sent to the SD UNet cross-attention described by Eq. (3) in Sec. 3.4 to further improve the texture quality.

3.4. Image Editing Pipelines

We exemplify our pipeline using reposing. Following the blue arrows in Fig. 2a, the denoising process is guided by a target pose, which will be enriched by textures from the source person. The texture information has two sources: the part features B in Sec. 3.1, and the pose-warped texture I_{tex} in Sec. 3.2. The part features B preserve style information, helping maintain the overall authenticity of the generated clothing, and I_{tex} provides detailed and spatial aligned textures, ensuring high fidelity in the generated image.

With B and I_{tex} serving as the texture sources, their information is transmitted by a cross-attention with the intermediate layers of SD UNet decoder:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T) \cdot V, \tag{2}$$

$$Q^i = W_Q^i h^i,\ K^i = W_K^i[B; E],\ V^i = W_V^i[B; E], \tag{3}$$

where h^i is the i^{th} intermediate feature representation of SD UNet decoder. W_Q^i, W_K^i and W_V^i are learnabled weights. E indicates the encoded pose-warped texture in the conditioning encoder. In the following, we use E_{rp}, E_{vt} and E_{tm} to denote the encoded pose-warped texture for each task.

Finally, with the SD denoising function f_θ, we obtain the latent code for reposing $I_{rp}^{(t)}$ at time step t by

$$I_{rp}^{(t)} = f_\theta\Big(g_\phi([I_{tex\text{-}rp}; P_{tgt}; I_{bg}]), B, E_{rp}, I_{rp}^{(t+1)}, y\Big),$$

where y is the optional text prompt that will also be mapped to the UNet decoder via the standard cross-attention block in SD [32]. This text cross-attention is applied after the part cross-attention in Eqs. (2) and (3).

In virtual try-on (pink dashed arrows in Fig. 2a), the source garment G_s is first removed and then replaced by the target garment G_t in the part features. Let $I_{src} - G_s$ be the image without the source garment. The part features in virtual try-on thus becomes $B' = [d_\omega(I_{src} - G_s); d_\omega(G_t)]$. B' is then utilized in denoising as

$$I_{vt}^{(t)} = f_\theta\Big(g_\phi([I_{tex\text{-}vt}; P_{src}; I_{bg}]), B', E_{vt}, I_{vt}^{(t+1)}, y\Big),$$

where the source pose P_{src} is used as guidance since virtual try-on doesn't change the original posture of the person.

Our model can also be used to edit the garment according to a text prompt (brown dashed arrows in Fig. 2a). Similar to virtual try-on, the described source garment G_s is removed from the source image I_{src}, for which we get B'. The garment's missing information will be replenished by the text cross-attention in SD, resulting in the following denoising process,

$$I_{tm}^{(t)} = f_\theta\Big(g_\phi([I_{tex\text{-}tm}; P_{src}; I_{bg}]), B', E_{tm}, I_{tm}^{(t+1)}, y\Big).$$

3.5. Objective Functions

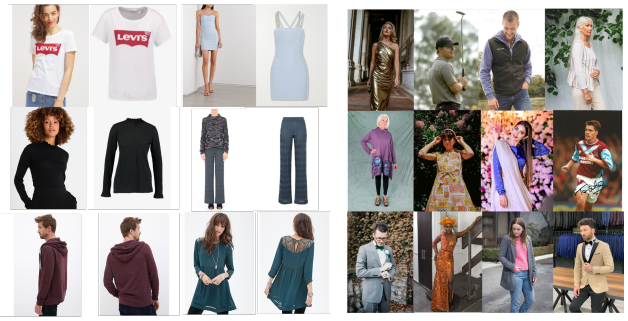
To prevent the texture blending problem [37], we apply two loss functions to constrain the cross attention for different human parts and pose-warped texture. For each human part p_n, let A_{p_n} and M_{p_n} be the attention map of B_{p_n} and segmentation map (resized to the same size), respectively. Following [37], we minimize their distance by:

$$\mathcal{L}_B = \sum\nolimits_n \Big(\text{mean}(A_{p_n} \odot (1 - M_{p_n})) - \text{mean}(A_{p_n} \odot M_{p_n})\Big).$$

Similarly, for the pose-warped texture, using the binary visibility map obtained from Sec. 3.2, we constrain the attention map of E by:

$$\mathcal{L}_E = \text{mean}(A_v \odot (1 - M_v)) - \text{mean}(A_v \odot M_v).$$

Driven by the two losses, our model is steered towards sampling from the pose-warped textures for visible pixels, and from part features for invisible regions. The net result is



(a) Existing datasets [8, 18, 24, 25]. (b) LH-400K.

Figure 3. Representative examples from different datasets. Our LH-400K includes people of diverse ages and backgrounds.

	Dataset	#Train Images	#Test Pairs	Background
In-Domain	DeepFashion [18]	44,096	8,512	Gray&White
	DressCode [25]	96,784	5,400	Walls
	VITON-HD [8]	23,294	2,032	Gray&White
	LH-400K (Ours)	409,270	-	Diverse
Out-Of-Domain	WPose (Ours)	-	2,304	Diverse
	WVTON (Ours)	-	440	Diverse

Table 1. Statistics for the training and evaluation sets used in our experiments. LH-400K provides large-scale diverse human data for training (Sec. 4.1), while WPose and WVTON are collected to evaluate the out-of-domain generalization capacity (Sec. 4.2).

a harmonious interplay that ensures accurate reconstruction and optimized fidelity for the entire generated content.

For standard SD training loss, in practice, the UNet predicts a noise ϵ_θ given the noisy version of the target image $I_{tgt}^{(t)}$ at each time step t. Let ϵ be the ground-truth noise and the L2 loss function is simplified as

$$\mathcal{L}_{SD} = \mathbb{E}_{I_{tgt},\epsilon\sim\mathcal{N}(0,1),t}\|\epsilon - \epsilon_\theta(I_{tgt}^{(t)}, t, ...)\|_2^2,$$

where ... omits other conditional inputs in our model, including B, P_{tgt} (or P_{src}), I_{bg}, I_{tex} and y. In summary, the overall objective function of our model is

$$\mathcal{L} = \mathcal{L}_{SD} + \lambda_1\mathcal{L}_B + \lambda_2\mathcal{L}_E,$$

where λ_1 and λ_2 are trade-off parameters.

4. Collecting High-Quality Human Images

Exiting datasets for human image editing exhibit limitations in scale and diversity since most of these images were in-studio photography with fashion models (*e.g.*,[8, 18, 24, 25]). This results in images with simple indoor backgrounds and younger age groups, as depicted in Fig. 3a. To address this, we meticulously curate a larger-scale training dataset with augmented diversity (Sec. 4.1). Additionally, we collect new challenging test sets for out-of-domain evaluation of existing models (Sec. 4.2).

4.1. Expanding the Training Data

We introduce LH-400K, a large-scale dataset of high-quality single-human images selected from LAION-400M [33] with diverse backgrounds, age groups, and body shapes. Notable distinctions between LH-400K and existing human image editing benchmarks are presented in Fig. 3 and Tab. 1. To build LH-400K, we follow several criteria and build an automated pipeline to filter clean human images from the large noisy pool. By combining the existing modest-scale paired data with our newly introduced large-scale unpaired data during training, our model can better adapt to diverse real-world scenarios. See Supp. for details of how we incorporate LH-400K in our model training.

The image quality criteria for LH-400K are: 1) image resolution is no less than 512 pixels; 2) image contains only one person; 3) head is visible; 4) pose can be detected by pose detection models; 5) little to no occlusion between the human and other objects in the scene; 6) clothing covers an adequate area of the human 7) image-text similarity is larger than 0.2. To fulfill these criteria, we applied human detector [30], face detector [27], pose detector [4, 14], instance segmentation model [6], human parser [22] and CLIP [29] on images from LAION-400M. Our resulting dataset contains 409,270 clean human images with diverse backgrounds.

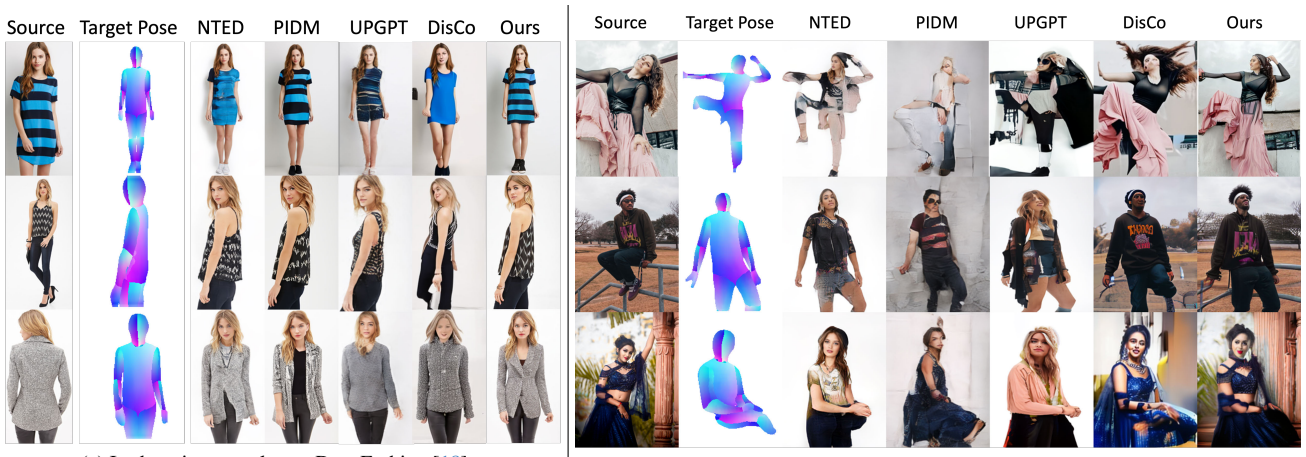
4.2. Test Data Collection

To evaluate the model's generalization capability to real-world unseen data, we collect two out-of-domain test sets: WPose for human reposing; and WVTON for virtual try-on. WPose comprises 2,304 real-world human image pairs with individuals adopting diverse postures (*i.e.*, dancing, squatting, lying down, *etc.*) that are rarely encountered in the training data. The backgrounds in WPose also encompass a diverse spectrum, ranging from indoor settings to outdoor scenes. Similarly, in WVTON, we collected 440 test pairs using garment images from Stock photos that comprise clothing items with diverse graphic patterns and fabric textures, to serve as our in-the-wild test set for virtual try-on. We confirmed our test data have a CC-0 or comparable license that allows academic use.

5. Experiments

Datasets. See Tab. 1 for dataset statistics. We train one model each for 256x256 and 512x512 images. As a unified model, **one** single trained model is used to evaluate on **all** tasks and on both in-domain and out-of-domain samples. See Supp. for our implementation details.

Metrics. We use Structural Similarity Index Measure (SSIM) [35], Frechet Inception Distance (FID) [15], Kernel Inception Distance (KID) [3], and Learned Perceptual Image Patch Similarity (LPIPS) [46] to evaluate image quality. On in-the-wild reposing test data, we use Masked-SSIM



(a) In-domain examples on DeepFashion [18].

(b) Out-of-domain examples on WPose.

Figure 4. Visualized results of reposing (256x256). Our model transfers the texture patterns better, particularly in out-of-domain samples. More results can be found in Supp.

	In-Domain DeepFashion [18]						Out-of-Domain WPose					
	256x256			512x512			256x256			512x512		
	FID↓	SSIM↑	LPIPS↓	FID↓	SSIM↑	LPIPS↓	FID↓	M-SSIM↑	M-LPIPS↓	FID↓	M-SSIM↑	M-LPIPS↓
NTED [31]	8.851	0.766	0.175	7.685	0.770	0.186	90.542	0.774	0.221	95.953	0.794	0.200
PIDM [2]	6.401	0.788	0.150	**5.899**	0.771	0.178	88.433	0.774	0.231	92.439	0.797	0.211
UPGPT [7]	9.611	0.762	0.185	9.698	0.765	0.184	75.653	0.784	0.202	85.798	0.795	0.196
DisCo [34]	9.842	0.763	0.184	9.818	0.770	0.201	50.948	0.795	0.181	58.331	0.813	0.180
UniHuman	**5.089**	**0.815**	**0.123**	6.092	**0.807**	**0.148**	**27.571**	**0.810**	**0.159**	**27.748**	**0.824**	**0.147**

Table 2. Quantitative results for reposing. Our UniHuman outperforms current methods on both in-domain and out-of-domain data.

Methods	PIDM [2]	UniHuman	DisCo [34]	UniHuman
Pose Accuracy	15.7%	**84.3%**	27.9%	**72.1%**
Texture Consistency	12.4%	**87.6%**	35.9%	**64.0%**
Face Identity	11.5%	**88.5%**	24.3%	**75.7%**

Table 3. Human evaluation results on WPose for reposing. Our UniHuman is preferred by users on all three evaluation methods.

and Masked-LPIPS computed on the target person region to exclude the irrelevant background from contributing to the metrics. All the compared methods use the same image size and padding when computing the above metrics.

5.1. Reposing Experiments

Quantitative Results. Tab. 2 compares our methods with NTED[31], PIDM [2] and UPGPT [7], DisCo [34]. NTED and PIDM are reposing methods trained on DeepFashion. UPGPT is a multi-task model trained on DeepFashion. DisCo is a reposing approach trained on combined datasets of 700K images, including DeepFashion, TiKTok dance videos [17], and *etc*. Since DisCo has not yet released its model at 512x512 resolution, we upsampled its generated 256x256 images and reported the numbers. In Tab. 2, PIDM has a slightly better FID score at 512x512 resolution on DeepFashion but worse scores on other metrics, suggesting that it generates more realistic texture yet fails to keep the identities of the original clothing. Overall, our model shows significant gains on most metrics, particularly on out-of-domain test data. Our model boosts the FID from 58 to 27 at 512x512 image resolution, demonstrating the better generalization capacity of the proposed method.

Qualitative Results. Fig. 4a shows our model's capacity at reconstructing intricate details. For example, note the equally distanced stripes in the first row and the accurate preservation of triangle patterns in the second row. For out-of-domain test samples in Fig. 4b, models trained on a single dataset (PIDM, NTED, and UPGPT) exhibit a recurring limitation. They transform the backgrounds, faces, and clothing to be similar to training samples, indicating poor generalization. Even though DisCo utilized more paired images from a video dataset to train their model, the last two rows of Fig. 4b show the resulting images have notable texture change after reposing. In contrast, our UniHuman preserves the clothing identity in synthesized images, *e.g.*, the graphic logo on the black t-shirt in the middle row.

User Studies. We randomly chose 200 images from the out-of-domain test data WPose, and each generated image was evaluated by three workers from Amazon MTurk to com-

(a) In-domain examples on DressCode [25].

(b) Out-of-domain examples on WVTON.

Figure 5. Virtual try-on results (512 x 512). Our UniHuman better recovers the intricate details in the target garment, particularly in out-of-domain samples. More results can be found in Supp.

	In-Domain												Out-of-Domain	
	VITON-HD [8]						DressCode [25]						WVTON	
	Paired				Unpaired		Paired				Unpaired		Unpaired	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓
PASTA-GAN [38]	-	-	-	-	-	-	-	-	-	-	-	-	179.138	6.561
HR-VITON [20]	7.533	0.160	0.902	0.075	9.979	0.200	-	-	-	-	-	-	159.553	3.269
GP-VTON [39]	5.572	0.069	**0.913**	**0.064**	8.641	0.589	9.111	0.584	0.900	0.080	10.600	1.205	-	-
LaDI-VTON [26]	5.647	0.047	0.901	0.070	**8.249**	0.078	4.820	0.106	0.920	0.059	6.727	0.182	147.375	2.416
UniHuman	**5.238**	**0.036**	0.905	0.068	8.312	**0.072**	**3.947**	**0.059**	**0.929**	**0.053**	**5.529**	**0.134**	**127.866**	**1.671**
GP-VTON(w.o. bg)			—-				5.649	0.234	0.930	0.043	7.067	0.311	-	
UniHuman(w.o. bg)			—-				**3.446**	**0.067**	**0.935**	**0.040**	**5.558**	**0.139**	-	

Table 4. Quantitative results for virtual try-on (512x512). KID is multiplied by 100. Our UniHuman, outperforms current methods on most metrics, particularly in out-of-domain data.

Methods	HR-VITON	UniHuman	LaDI-VTON	UniHuman
Texture Consistency	23.6%	**76.4%**	24.1%	**75.9%**
Image Fidelity	21.6%	**78.4%**	27.7%	**72.3%**

Table 5. User studies on out-of-domain test data for the virtual try-on task. HR-VITON [20] and LaDI-VTON [26] are virtual try-on methods. Our model is preferred in at least 72% cases.

pare with the synthesized image of an existing method (*i.e.*, PIDM or DisCo). These workers evaluate the image quality on three aspects: pose accuracy, clothing texture consistency, and face identity consistency. Tab. 3 reports the study results, where our method is preferred in all three aspects.

5.2. Virtual Try-on Experiments

Results. Tab. 4 reports our results with state-of-the-art performance over other methods specifically designed for virtual try-on. Note that the original GP-VTON [39] removed all backgrounds in DressCode, and in this setting, we still outperform them (the last two lines of Tab. 4). Some of our gains are likely attributed to the reposing data we can take advantage of due to using a unified model. Specifically, there are only 120,078 virtual try-on data samples in our training set, but 453,366 reposing samples. Additionally, compared with try-on-only models, we also generalize better to out-of-domain samples, where our FID score improves from 147 to 128. As shown in Fig. 5, our model successfully reconstructs the texture patterns of unseen garments for a variety of body shapes.

User Studies. Similar to the human evaluation in Sec. 5.1, we conducted user studies for virtual try-on. This time, we asked the workers to evaluate the image fidelity and the texture consistency of the try-on garment, respectively. Tab. 5 shows that our model is preferred in at least 72% of test cases, demonstrating the superior performance of the proposed method for in-the-wild scenarios.

		In-Domain							Out-of-Domain				
		DeepFashion [18]			VITON-HD [8]		DressCode [25]		WPose			WVTON	
		FID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	FID↓	M-SSIM↑	M-LPIPS↓	FID↓	KID↓
(a)	w.o. I_{tex}	6.016	0.798	0.143	9.953	0.295	7.042	0.279	31.352	0.797	0.178	146.352	2.134
	w.o. L_E	5.232	0.802	0.131	9.584	0.185	6.981	0.275	30.267	0.802	0.169	140.187	2.036
(b)	w.o. 400K & I_{tex}	6.155	0.813	0.123	9.998	0.316	7.210	0.280	47.736	0.791	0.178	165.736	3.120
	w.o. 400K	5.269	**0.826**	**0.110**	9.659	**0.184**	6.960	0.271	37.826	0.804	0.167	142.839	2.139
	rp only	5.682	0.813	0.124	17.345	1.005	14.909	0.867	42.587	0.790	0.185	170.322	3.180
	UniHuman	**5.089**	0.815	0.123	**9.558**	0.248	**6.310**	**0.208**	**27.571**	**0.810**	**0.159**	**131.500**	**1.730**

Table 6. Ablation results. Image resolution is 256x256. KID is multiplied by 100. Our full model achieves the best performance.

Figure 6. Results of text manipulation. Our model manipulates the clothing to match the specified concept.

5.3. Text Manipulation Experiments

The proposed method can manipulate the clothing shape, category, and texture by text description, as showcased in Fig. 6. We chose UPGPT [7] and EditAnything [11] as our baselines since both methods have been used for fashion item editing. Notably, our model manages to maintain the original pose while manipulating the image content to match the specified concept. For example, both EditAnything and our model turn the lady's clothing into a Halloween costume on the second row, whereas only our method keeps the original body pose unchanged. See Supp. for more analysis about text manipulation.

5.4. Ablations

Pose-Warped Texture. In Tab. 6(a), **w.o.** $\boldsymbol{I_{tex}}$ removes the pose-warped texture. Consequently, we notice a marked drop in performance across all metrics, underscoring the critical role played by the pose-warped texture and the significance of visible pixels it provides. In a separate experiment, **w.o.** $\boldsymbol{L_E}$ includes the pose warped-texture but removes the loss function L_E, which constrains the attention area of the pose-warped texture. This model shows better performance than **w.o.** $\boldsymbol{I_{tex}}$, but it still has a significant gap with our full model. By introducing the pose-warping module and L_E, our full model yields the most substantial gains across both in-domain and out-of-domain test sets.

Data. To validate the effectiveness of the collected LH-400K, we train an ablation model with only existing datasets, denoted by **w.o. 400K** in Tab. 6(b). Removing LH-400K results in comparable or even superior in-domain performance (*i.e.*, better SSIM and LPIPS). However, out-of-domain metrics suffer from a significant drop, indicating overfitting and poor generalization due to limited data scale and diversity. In **rp only**, we exclude both LH-400K and try-on datasets, relying solely on DeepFashion for training the reposing task. The performance on both reposing datasets, *i.e.*, in-domain DeepFashion and out-of-domain WPose, has a notable decline compared to **w.o. 400K**. This demonstrates the beneficial impact of the try-on data on the reposing task. Additionally, further removing I_{tex} (**w.o. 400K &** $\boldsymbol{I_{tex}}$) causes a more pronounced drop, emphasizing the importance of the pose-warped texture. More ablation experiments can be found in Supp.

6. Conclusion

We propose UniHuman, a unified model that achieves high-quality human image editing results on multiple tasks for both in-domain and out-of-domain data. We leverage human visual encoders and a lightweight pose-warping module that exploits different pose representations to accommodate unseen textures and patterns. Furthermore, to expand the diversity and scale of existing datasets, we curated 400K high-quality image-text pairs for training and collected 2K human image pairs for out-of-domain testing. Experiments on both in-domain and out-of-domain test sets demonstrate that UniHuman outperforms task-specific models by a significant margin, earning user preferences in an average of 77% of cases. In future work, we will explore the adaptation of UniHuman to the video domain.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional styleGAN. *ACM Transactions on Graphics*, 40(6):1–11, 2021. 2, 3

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 6

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations*, 2018. 5

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the International Conference on Machine Learning*, 2023. 1, 2

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 5

[7] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. UPGPT: Universal diffusion model for person image generation, editing and pose transfer. In *Workshop on the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 6, 8

[8] 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 conference on computer vision and pattern recognition*, 2021. 2, 5, 7, 8

[9] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 1

[10] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[11] Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. EditAnything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the ACM International Conference on Multimedia, Demo track*, 2023. 1, 8

[12] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 1, 2

[13] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2018. 5

[15] 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. In *Advances in Neural Information Processing Systems*, 2017. 5

[16] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[17] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6

[18] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2Human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2, 5, 6, 8

[19] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 2

[20] 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*, 2022. 2, 3, 7

[21] Nannan Li, Kevin J Shih, and Bryan A Plummer. Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[22] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 5

[23] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016. 2, 5

[25] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5, 7, 8

[26] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 1, 2, 7

[27] Ravi Teja Mullapudi, William R Mark, Noam Shazeer, and Kayvon Fatahalian. Hydranets: Specialized dynamic architectures for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 5

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[29] 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 *Proceedings of the International Conference on Machine Learning*, 2021. 2, 5

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 5

[31] 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*, 2022. 6

[32] 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*, 2022. 2, 3, 4

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 5

[34] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. DisCo: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2, 6

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[36] Zijian Wang, Xingqun Qi, Kun Yuan, and Muyi Sun. Self-supervised correlation mining network for person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 2

[37] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 4

[38] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. In *Advances in Neural Information Processing Systems*, 2021. 2, 7

[39] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 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*, 2023. 2, 3, 7

[40] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[41] 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 conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[42] Hanrong Ye and Dan Xu. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *Proceedings of the International Conference on Learning Representations*, 2022. 1

[43] 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*, 2021. 1

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4

[45] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2018. 5

[47] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[48] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. TryOnDiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2