CLOTHFIT: CLOTH-HUMAN-ATTRIBUTE GUIDED VIRTUAL TRY-ON NETWORK USING 3D SIMULATED DATASET

Yunmin Cho ^{1,2}, Lala Shakti Swarup Ray ^{1,2}, Kundan Sai Prabhu Thota ^{2,3}, Sungho Suh^{1,2*}, Paul Lukowicz^{1,2}

Department of Computer Science, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany
 German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
 Shapematchr GmbH, Berlin, Germany

ABSTRACT

Online clothing shopping has become increasingly popular, but the high rate of returns due to size and fit issues has remained a major challenge. To address this problem, virtual try-on systems have been developed to provide customers with a more realistic and personalized way to try on clothing. In this paper, we propose a novel virtual try-on method called ClothFit, which can predict the draping shape of a garment on a target body based on the actual size of the garment and human attributes. Unlike existing try-on models, ClothFit considers the actual body proportions of the person and available cloth sizes for clothing virtualization, making it more appropriate for current online apparel outlets. The proposed method utilizes a U-Net-based network architecture that incorporates cloth and human attributes to guide the realistic virtual try-on synthesis. Specifically, we extract features from a cloth image using an auto-encoder and combine them with features from the user's height, weight, and cloth size. The features are concatenated with the features from the U-Net encoder, and the U-Net decoder synthesizes the final virtual try-on image. Our experimental results demonstrate that ClothFit can significantly improve the existing state-of-the-art methods in terms of photo-realistic virtual try-on results.

Index Terms— Virtual try-on, Cloth simulation, Garment fitting, Digital apparel

1. INTRODUCTION

Image-based virtual try-on has become an increasingly popular method for customers to purchase clothing online, as it allows them to visualize how a particular garment would look on their body before making a purchase. In the e-commerce clothing market, customers often encounter issues with clothing size and fit, which leads to a high rate of returns. According to recent reports [1, 2], the biggest challenges related to apparel e-commerce non the standardized use of size charts

by retailers, which makes it complicated for customers to find the proper fit and different body shapes and individual body proportions makes it challenging to visualize how the clothing would look on the customer. These challenges have led to a growing interest in virtual try-on systems, which provide customers with a more realistic and personalized way to try on clothing. However, accurately simulating the fit of a garment on a customer's unique body shape is still a challenging problem. Existing virtual try-on methods [3, 4, 5, 6, 7] have focused on realistic visualization of clothed human images, but have not adequately addressed the issue of garment fit.

One of the main motivations of our work is to provide a more accurate and realistic virtual try-on experience for users. The ability to simulate the fit of a garment on a user's body can help reduce the number of returns and increase customer satisfaction, ultimately benefiting both customers and online retailers. Recently, many virtual try-on methods have been proposed to utilize deep learning techniques to generate realistic clothing images. For example, Han et al. [3] proposed a multitask encoder-decoder network, called VITON, to warp the target garment on the target human clothed area without using any 3D information. Similarly, Fele et al. [7] proposed a context-driven virtual try-on network (C-VTON) to align the target clothing to the segmented body parts with geometric matching. However, many of these methods still suffer from limitations when it comes to accurately simulating the fit of a garment on a user's body. Some methods rely on pre-defined garment templates, which may not accurately capture the complexity of real-world garments. Other methods may require extensive user input, such as 3D scans of the user's body, which can be time-consuming and expensive. Our earlier work [8] proposed an estimation method of 3D body shape and clothing measurements from frontal and side view images to address the size issue.

In this paper, we propose a novel virtual try-on method that utilizes a U-Net-based [9] network architecture to estimate the fit of a garment on a user's body. As an extension to body shape estimation from RGB images, this method predicts the draping shape of the garment on the estimated 3D body shape. The proposed method takes as input a frontal im-

Corresponding author: sungho.suh@dfki.de

This work was supported by the BMBF (German Federal Ministry of Education and Research) in the VidGenSense (01IW21003). The Carl-Zeiss Stiftung also funded it under the Sustainable Embedded AI (P2021-02-009).

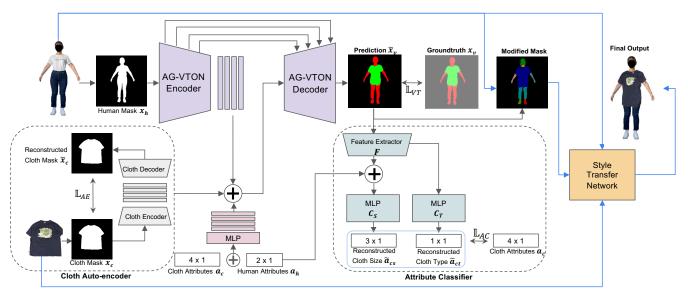


Fig. 1: ClothFit architecture: An auto-encoder extracts the features from a cloth mask and an attribute classifier estimates cloth attributes. A U-Net-based attribute-guided virtual try-on network (AG-VTON) generates a masked image of a clothed human using input cloth image, cloth and human attributes, and a frontal-view image of the user. A style transfer network with a modified segmented mask generates the final output image of the user with the input cloth.

age of the user with height and weight information, as well as an image of the clothing they wish to purchase along with its size factors. We utilize an auto-encoder network to extract the feature vector of the clothing image, and an MLPbased feature extractor to extract the features of the garment attributes and user factors. We then concatenate these features with the features from the U-Net encoder and use the U-Net decoder to synthesize a virtual try-on image. To train our proposed networks, we generate synthetic images simulated over 3,801 people with four different cloth types using the 3D physics simulator Blender [10] to overcome the limitation of collecting various cloth types, sizes, and human body shapes. Our proposed method is a cloth-humanattribute-guided virtual try-on network (AG-VTON), and we utilize a style transfer network to generate a photo-realistic try-on image. The experimental results demonstrate that the proposed method can synthesize the photo-realistic results by changing the attributes of the user and the target garment, and provides more realistic results than other state-of-the-art methods, which does not show any difference between two different sizes of the cloth.

The main contributions of our study can be summarized as follows: (1) A large synthetic dataset with frontal images of 3801 body types with different height and weight attributes wearing 45612 cloth types along with cloth attributes. (2) A novel virtual try-on method, called ClothFit, to predict the draping shape of the garment over a target body based on the actual size of the garment and human attributes. (3) Validation of the proposed framework with both synthetic and real-world datasets.

The remainder of this paper is organized as follows. Sec-

tion 2 provides an overview of the network architecture and training process. Section 3 presents our experimental results and analysis. Finally, in Section 4, we conclude our work and give an insight into future work.

2. PROPOSED METHOD

The proposed framework consists of four networks: an autoencoder for extracting features of a target clothing image, an attribute classifier for estimating the actual size of the clothing from a clothed human image, a U-Net-based attribute-guided virtual try-on network (AG-VTON), and a style transfer network incorporating the cloth mask and the input user image. The overview of the proposed framework is shown in Fig. 1, where the input is the user image and cloth image of the size of 512 x 512, and the output the clothed human image.

2.1. Generating Synthetic Data using 3D Simulation

To effectively train the proposed cloth-human-attribute-guided virtual try-on networks, we require a large dataset that includes human images, cloth images, and attributes such as human factors and cloth sizes. However, no existing virtual try-on dataset has all of these components. To address this issue, we generate synthetic images using the 3D physics simulator, Blender [10]. Our dataset comprises three types of images: a target clothing image, a frontal-view image of the user, and a clothed human image. We also include two types of attributes: cloth attributes, including cloth type, chest circumference, total length, and sleeve length, and human attributes, including height and weight of the user. To generate the dataset, we first created 3D SMPL bodies [11] of vary-

ing heights and weights using the SMPL-X Blender add-on. Next, we designed cloth sewing patterns for each type of cloth based on their actual design. The sewing patterns were used to join together the different cloth pieces, which were then deformed over the SMPL body mesh to produce the clothed human images [12]. To facilitate differentiation between the body and cloth, different textures were assigned to the body and cloth. The garment sewing pattern's size was modified to reflect different sizes of the garment. We automated the entire process of rendering cloth and collecting attribute data using Blender API, enabling us to generate a large dataset with a wide range of cloth sizes and body shapes.

2.2. Attribute-Guided Virtual Try-On

In the proposed network architecture, we train three networks: a cloth auto-encoder, a U-Net-based AG-VTON, and an attribute classifier. First, the cloth auto-encoder and attribute classifier are trained separately. The cloth auto-encoder takes a cloth mask image of size 512×512 as input and extracts a feature vector of size $\frac{W}{16} \times \frac{H}{16} \times 256$. The encoder comprises four convolution and max-pooling layers, and the decoder contains four upsampling and convolution layers. The loss function of the cloth auto-encoder is defined as follows.

$$\mathbb{L}_{AE} = \mathbb{L}_{BCE}(\tilde{x}_c, x_c)$$
where
$$\mathbb{L}_{BCE} = \mathbb{E}_{p,q}[q \log p + (1-q) \log(1-p)],$$
(1)

where x_c and \tilde{x}_c denote the cloth mask image and reconstructed image, respectively, and \mathbb{L}_{BCE} denotes the standard binary cross-entropy loss (BCE).

The attribute classifier is used to classify the type of cloth and estimate the size of the cloth in the clothed human image. At the first stage, the attribute classifier is trained with the ground truth of the clothed human images, and at the second stage, the pretrained attribute classifier is used to improve the performance of the AG-VTON. The attribute classifier takes a clothed human mask image of size $512 \times 512 \times 2$ and the human attributes as input and outputs the cloth attributes as depicted in Fig. 1. The attribute classifier contains a feature extractor from ResNet-18 [13] and two multilayer perceptrons (MLP) for cloth type classification and cloth size estimation. The loss function of the attribute classifier is expressed as follows:

$$\mathbb{L}_{AC} = \mathbb{L}_{CE}(a_{ct}, \tilde{a}_{ct}) + \lambda \, \mathbb{L}_{MAE}(a_{cs}, \tilde{a}_{cs})$$
where $\tilde{a}_{ct} = C_T(F(x_v)), \quad \tilde{a}_{cs} = C_S(F(x_v), a_h)$ (2)

and x_v denotes the clothed human mask image, a_{ct} , a_{cs} , and a_h denotes the cloth type, the cloth sizes in the cloth attribute, and the human attribute, respectively, F, C_T , C_S are the feature extractor, cloth type classifier, and cloth size estimator, respectively, and \mathbb{L}_{CE} and \mathbb{L}_{MAE} denote the standard crossentropy (CE) and mean absolute error (MAE) losses.

In the second stage of the proposed method, the U-Netbased AG-VTON is trained using the cloth auto-encoder and attribute classifier trained in the previous step. The virtual try-on network takes as input a silhouette image of the user, a target cloth mask image, and cloth and human attributes. An MLP-based feature extractor transforms the attribute vector, which is then reshaped and concatenated with the AG-VTON encoder's output and the feature vector extracted from the cloth auto-encoder. This controls the cloth area on the target human body. The synthesized virtual try-on mask is then fed into the pretrained attribute classifier to classify the cloth type and estimate cloth sizes. Thus, the virtual try-on network is trained to synthesize the clothed human mask to minimize the loss between the output of the network and ground truth, and the loss between the input cloth attributes and the cloth attributes estimated by the attribute classifier. Finally, the objective functions of the proposed method are defined as follows:

$$\mathbb{L}_{VT} = \alpha \, \mathbb{L}_{DICE}(x_v, \tilde{x}_v) + \beta \, \mathbb{L}_{CE}(a_{ct}, \tilde{a}_{ct}) + \gamma \, \mathbb{L}_{MAE}(a_{cs}, \tilde{a}_{cs})$$
where
$$\mathbb{L}_{Dice}(p, q) = 1 - 2 \frac{\sum_i p^i q^i}{\sum_i p^i \sum_i q^i}, \quad \tilde{x}_v = G(x_h, x_c, a_c, a_h)$$
(3)

and G is the AG-VTON, x_h and x_c denote the input user mask and the cloth mask, respectively, a_c and a_h are the cloth attribute and human attribute, respectively, and \mathbb{L}_{DICE} denotes the dice loss, which is adopted for complex and highly imbalanced datasets to distinct properties of foreground and background.

2.3. Style Transfer

To generate realistic images of clothed human, we utilize a style transfer network since our proposed networks generate masks rather than texturized cloth. We use a modified version of SieveNet [14] for converting our cloth mask into textured cloth. We replace the conditional segmentation mask generation module with our own modified version of the segmented mask, which is generated by superimposing the cloth mask from our model over the segmented mask of the whole body of the person generated using self-correction human parsing [15]. The modified segmented mask is used with a coarse-to-fine warping module and a segmentation-assisted texture translation module to generate the person wearing the cloth.

3. EXPERIMENTAL RESULTS

Dataset: The cloth dataset used in this work consists of four different types of clothes: t-shirt, long-sleeve, dress, and blazer, as illustrated in Fig. 2. For each type, we predefined the minimum and maximum size values. Random sizes were generated within these ranges, resulting in a diverse dataset containing various types and sizes of clothes. As depicted in Fig. 2, the dataset contains clothes with different sizes, and the draping shape of each cloth varies depending on its size. Specifically, we generated 8,412 images of clothed females for t-shirts, long-sleeved shirts, and dresses, and 6,792 images of clothed males for t-shirts, long-sleeved shirts, and

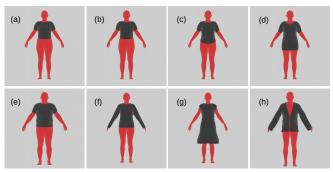


Fig. 2: Example images of four size variations for the same t-shirt: (a) S, (b) M, (c) L, and (d) XL. Examples of four cloth types: (e) t-shirt, (f) long sleeve, (g) dress, and (h) blazer.

Table 1: Quantitative evaluation of the AG-VTON and ablation result of attribute classifier

	F1 ↑	IoU ↑	J↑ Average length error (pixel) ↓	
w/ AC	0.9436	0.8933	11.7589	
w/o AC	0.9409	0.8886	11.8511	

blazers. In total, our proposed network was trained with 45,612 images.

Implementation Details: For the evaluation of the virtual try-on mask image synthesis, we used F1, IoU scores, and average length error in pixel. To compare the proposed method with the state-of-the-art methods quantitatively, we use Frechet Inception Distances (FID) [16] and Learned Perceptual Image Patch Similarities (LPIPS) [17]. The proposed networks were trained for 300 epochs with a batch size of 24, using Adam optimizer with a 0.001 learning rate. We implemented the proposed networks using PyTorch in NVIDIA GeForce RTX 3090.

Results on Synthetic Dataset: We evaluated the performance of the proposed method on a data that is not used for training process. Our method achieved high F1 and IoU scores, as reported in Table 1. We also conducted ablation experiments to investigate the contribution of the attribute classifier. We found that training the network with the attribute classifier led to higher F1 and IoU scores, indicating improved accuracy in virtual try-on mask image synthesis. Additionally, the attribute classifier helped to predict accurate cloth sizes, as reflected in the lower average length-pixel difference. Specifically, we measured the number of pixels in the cloth length and observed that the attribute classifier effectively controlled the cloth length.

Results on Real Dataset: Fig. 3 presents the qualitative evaluation with a real cloth and human image along with inference from other state-of-the-art models, such as C-VTON [7] and Flow-Style-VTON [18]. Current 2D virtual try-on methods follow a two-step process of image wrapping(estimation of target cloth shape) followed by texture transfer (estimation of target cloth texture). To compare our model with previous state-of-the-art (SOTA) models, we replaced the first step of SieveNet with AG-VTON output and computed the FID and LPIPS score that measures the simi-

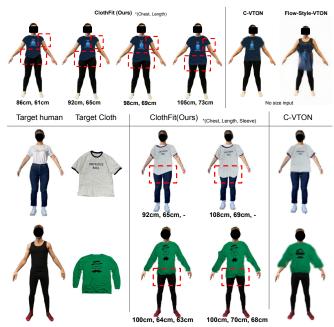


Fig. 3: Result on real dataset: The same human wearing the same cloth with different sizes using our method and comparison result with state-of-the-art on the same dataset.

Table 2: Quantitative comparison with the state-of-the-art method.

	C-VTON [7]	Flow-Style-VTON [18]	Clothfit (Ours)
FID↓	19.54	33.87	40.76
LPIPS ↓	0.107	0.210	0.128

larity of texture alignment between the input and synthesized image. As shown in Table 2, our proposed method generates outputs of comparable quality to current SOTA methods and even achieves a lower LPIPS score than Flow-Style-VTON, although our model is not explicitly trained for texture translation.

4. CONCLUSION

In this work, we proposed a novel virtual try-on system, ClothFit, that generates photorealistic images of a person wearing a garment based on the actual size of the garment and human attributes. The system consists of a cloth auto-encoder, an attribute classifier, and a U-Net-based AG-VTON. We trained the proposed networks on the dataset generated using 3D physics simulation, Blender. The experimental results showed that the proposed method synthesized the virtual try-on images better than state-of-the-art models and could generate the virtual try-on images along with the actual human and cloth attributes. Our proposed system has the potential to be used in various industries such as e-commerce and fashion design, where the virtual try-on system can assist customers to try on garments virtually before purchasing, saving time and resources.

5. REFERENCES

- [1] Barclay, "Return to sender: Retailers face a 'phantom economy' of £bn each year as shopper returns continue to rise," 2018.
- [2] 3DLook, "Counting the cost of fashion ecommerce's unsustainable apparel return rates," 2023.
- [3] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis, "Viton: An image-based virtual tryon network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7543–7552.
- [4] 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)*, 2018, pp. 589–604.
- [5] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert, "Image based virtual try-on network from unpaired data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5184–5193.
- [6] 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*, 2020, pp. 7850–7859.
- [7] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc, "C-vton: Context-driven image-based virtual tryon network," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3144–3153.
- [8] Kundan Sai Prabhu Thota, Sungho Suh, Bo Zhou, and Paul Lukowicz, "Estimation of 3d body shape and clothing measurements from frontal-and side-view images," in 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 2631–2635.
- [9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and

- Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
- [10] Blender Online Community, *Blender a 3D modelling* and rendering package, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.
- [11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.
- [12] Vjaceslav Tissen, SimplyClothPro.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [14] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai, "Sievenet: A unified framework for robust image-based virtual try-on," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 2182–2190.
- [15] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2020.
- [16] 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*, vol. 30, 2017.
- [17] 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 conference on computer vision and pattern recognition, 2018, pp. 586–595.
- [18] Sen He, Yi-Zhe Song, and Tao Xiang, "Style-based global appearance flow for virtual try-on," in *CVPR*, 2022.