TritonVITON : 3D-aware Real time Video Virtual Trial On for Immersive Online Fashion Fitting

Eric Zhao[1*], Huaijing Hong[1*], Jesus Gonzales[1*], Thanh-Long Nguyen Trong[1*], Bhavik Chandna[1†]

[1] University of California, San Diego, * Equal contribution † Team lead

SAIRS 2025
INNOVATION

Motivation & Background

- Digital fashion is a growing industry as e-commerce continues to redefine retail, resulting in more and more transactions to be made over the internet. This growing climate shows the need for a virtual try on, allowing users to virtually try on clothing that they are buying online to save them the trouble of having to return items online. **By 2030, the global virtual try-on market is projected to reach $15 billion**, driven by AI advancements and surging demand for immersive shopping experiences.

- Despite advances in virtual try-on technologies, most existing solutions remain locked behind steep paywalls, proprietary tools, or hardware-intensive systems. Prior efforts by startups like Zeekit and tech giants (e.g., Amazon's Echo Look, Meta's AR tools) have laid groundwork, yet adoption remains limited due to cost and technical barriers.

- Today, **71% of consumers** expect personalized interactions, but legacy systems fail to meet these expectations at scale. For example, Amazon, the biggest e-commerce platform in the world, only offers users a virtual try on service for some shoes, glasses, and experimentally some garments.

- **Introducing TritonVITON,** disrupting this landscape by offering real-time, webcam-based virtual fitting via diffusion models and pose estimation. This aligns with trends showing **80% of businesses** report **38% higher consumer spending** when experiences are personalized. We allow users to visualize how clothes look and move on their own bodies using nothing more than their standard webcam.

- **TritonVITON warps upper, lower, and full body garments** onto live video feeds while preserving key visual features like fabric texture, garment structure, and motion consistency. With support for **11 garment categories** and **latency under 2 seconds**, TritonVITON bridges accessibility gaps while meeting the **76% of users** who demand seamless, frustration-free shopping.

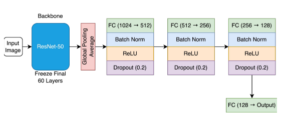
Cloth Category Classifier

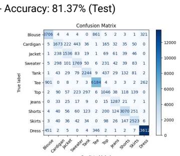
Transfer Learning:
- Leveraged a ResNet-50 pretrained on ImageNet.
- Fine-tuned the last deeper layers for garment classification while freezing early layers to retain general features.

Class Imbalance:
- Utilized weighted cross-entropy to prevent bias towards majority classes.
- Applied targeted data augmentation (e.g., flipping, color jitter) to boost diversity in minority classes.

Dataset & Split:
- Trained on DeepFashion-C containing over 240k richly annotated images with upper, lower, and full body garments. (Train 60%, Val 20%, Test 20%)

Results:
- Accuracy: 81.37% (Test)





Structural Guidance via Pose and Segmentation

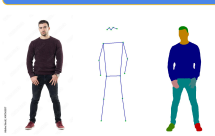
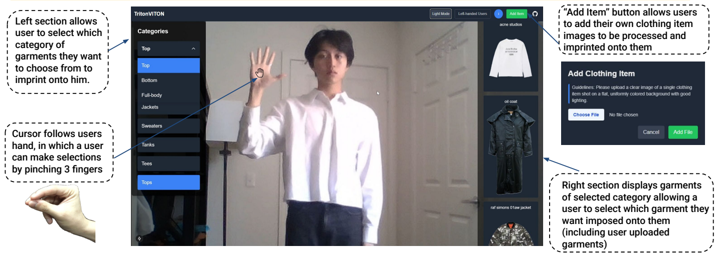


Table 1: Pose Estimation Models Comparison

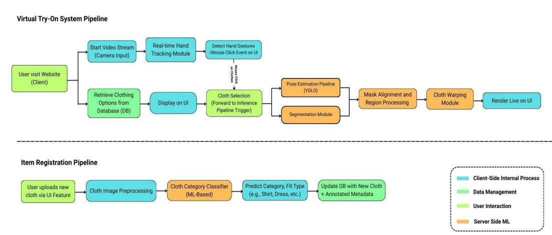
Model	FPS	Accuracy	Number of Landmarks
Mediapipe	15	low-medium	32
Yolov11	2-15	medium-high	17
OpenPose	7	high	17
VitPose	2	high	17

- Structural guidance is derived from pose estimation and segmentation maps, providing essential priors for accurate cloth warping and 3D-aware alignment.
- **YOLOv11** is used for accurate and fast body pose detection, offering modular variants via Ultralytics to balance inference speed and precision. You can refer to the table for comparison with different models to showcase the power of the model for our use case.
- **Self-Correction-Human-Parsing** is employed for segmentation due to its lightweight architecture, enabling efficient generation of person-specific masks in real time.
- The combination ensures that cloth alignment respects both human pose and body structure, enhancing realism in virtual try-on scenarios.

GUI Layout



Method Overview



Hand Tracking: Uses Mediapipe to detect hand gestures for UI control and simulate clicking.

Pose + Segmentation For Masking and Depth Cues: **Yolo v11** detects 3D body key points to guide cloth placement.
Self-Correction-Human-Parsing isolate the user for accurate overlay.

3D Clothes : Detailed 3D garments with texture and shape, which can be used to further enhance our cloth draping and physics. (Currently under development).

Cloth Warping : Warping model deform 2D clothes to fit the user realistically, preserving texture and depth cues.

Cloth Warping

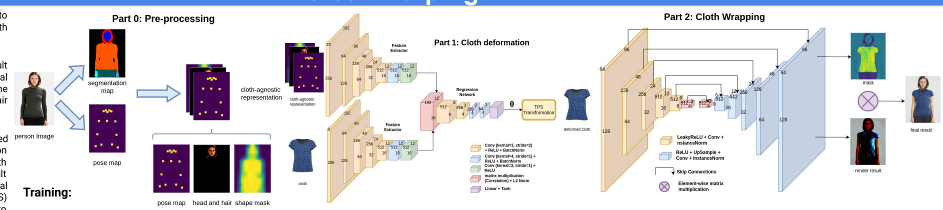
The general idea of this part is when given a cloth c and a person image *I*. We want to generate a result *I'* which preserve the pose from I while have the person wear given cloth *C*.This part can be divided into three part:

- Part 0: Pre-processing. In this part, the person image *I* is being processed and result with a cloth-agonistic representation *P* which was designed to preserved all the personal features of the person while eliminate all the information about the clothes. The cloth-agonistic representation *P* includes pose map, body shape mask, and Face and hair segment. It was done by body part segmentation model and pose estimation model.

- Part 1: Cloth deformation. In this part, the cloth-agonistic representation *P* is being used to deform given cloth *C*. We first use feature extractor module consist of 5 convolution layers to extract abstract features from both cloth-agonistic representation *P* and cloth image *C*. Then correlate there features by matrix multiplication and pass the result through another regression network (4 convolution layers) which output a spatial transformation parameters θ. By passing the parameters θ into a Thin-Plate Spline (TPS) transformation module *T*, we can generate a grid eventually deform the cloth *C* into wrapped cloth *C'*.

- Part 2: Cloth wrapping. In this part, a Unet shape network (6 down sample modules, 6 up sample modules and skip connections) is being used to eventually generate a rendered image I_r and a mask *M* with input being wrapped cloth *C'* and cloth-agonistic representation *P*. Our final output *I'* of the whole network is being given by (⊙ stands for pixel-wise matrix multiplication):

$$I' = M \odot C' + (1 - M) \odot I_r$$



Training:

Ideally, the training data should be image of person I (wearing cloth C1), image of person I (wearing cloth 2) and image of cloth 2 as input. However it is really hard to find this kind of dataset. Therefore, we use image of person I (wearing cloth C) and image of cloth C as input. By eliminating cloth information in part 0 make the training more generalizable and can have the same effect as ideal case.

- For part 1, the loss is L1 pixel-wise loss, defined as:

$$\mathcal{L}_{\text{part 1}} = ||C - C_t||_1 = ||T_\theta(C) - C_t||_1$$

C_t is the ground truth of cloth *C* being wore on person *I*, and *C'* is the deformed cloth.For part 2, the loss is a combination of perceptual loss and L1 loss, defined as:

$$\mathcal{L}_{\text{part 2}} = \lambda_{L1}||I - I_t||_1 + \lambda_{vgg}\mathcal{L}_{VGG}(I', I) + \lambda_{mask}||1 - M||_1$$

I_t is the ground truth image, *I'* and *M* is the result and mask generated by our model. The L_{vgg} is the perceptual loss which used pretrained VGG19 network pre-trained on ImageNet and defined as:

$$\mathcal{L}_{VGG}(I', I_t) = \sum_{i=1}^{5} \lambda_i ||\phi_i(I') - \phi_i(I_t)||_1$$

ϕ_i (I) denotes the feature map of image I of the i-th layer in the visual VGG19 network.

Testing:

In the next section you can see some testing results. Here is a chart which demonstrates the comparison between our model and other model on multiple different datasets.

Method	SSIM ↑	LPIPS ↓	$VFID_{I3D}$ ↓
VTON	0.462	0.545	6.371
FW-GAN	0.665	0.256	8.009
PBAFN	0.890	0.147	4.536
ClothFormer	0.924	0.101	3.767
Ours	**0.939**	**0.078**	**3.605**

Table 1: Quantitative comparison of different methods on virtual try-on.

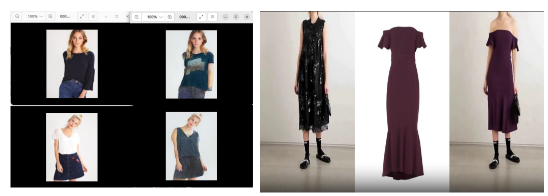
3D garments

To enhance the realism and adaptability of user-centric visual effects, we considered integrating 3D assets to facilitate better rendering. This approach builds upon advancements in depth-estimation algorithms and image-based rendering techniques to achieve geometrically consistent 3D reconstructions. A comparative evaluation of two state-of-the-art tools—VFusion3D and InstantMesh—revealed that InstantMesh achieves comparable visual fidelity while significantly reducing output file sizes.

Table 2: Image To Mesh File Size Comparison (in MB)

Clothing	VFusion3D	InstantMesh
Blouse	37.3	3.6
Dress	25.4	3.5
Pants	17	2.4

Results



Some examples for our Virtual Trial ON System - For Shirt and Dress.

Future Works

- Implementing a login system with user-specific accounts, as the current version lacks any user-based database.
- Incorporating 3D clothing assets in training to improve the physical realism and visual fidelity of the clothing, especially during large user motions in front of the camera.
- Adding the ability to capture videos or screenshots while wearing virtual clothes.
- Supporting simultaneous trials of multiple clothing items—currently, the system only allows one item at a time.

References