

3D Virtual Fitting Network (3D VFN)

Danyal Mahmood, Wei Wen Leong, Humaira Nisar^{id}^a and Ahmad Uzair bin Mazlan
*Department of Electronic Engineering, Faculty of Engineering and Green Technology,
Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia*

Keywords: Virtual Try-On, Virtual Fitting Room, Deep Generative Models, Geometric Matching, Depth Estimation.

Abstract: With the rise in digital technology and the fast pace of life, as well as the change in lifestyle due to the pandemic, people have started adopting online shopping in the garment industry as well. Hence, research on Virtual Try-On (VTO) technologies to be implemented in virtual fitting rooms (VFRs) has drawn significant attention. The existing VFR technologies rely on deep generative models with an end-to-end pipeline, from feature extraction to garment warping and refinement. While currently there are 2D and 3D VTO solutions, the 3D ones have enormous commercial potential in the fashion market as the technology has been proven effective for providing a photo-realistic and detailed try-on result. However, the existing 3D VTO solutions principally rely on annotated human body shapes or avatars, which are unrealistic. By integrating the technologies embedded in both 2D and 3D VTO solutions, this paper proposes a VTO solution that relies on geometric settings in the 3D space namely the 3D Virtual Fitting Network (3D VFN), that solely relies on 2D RGB garment and single-person human images as inputs, generating a photo-realistic warped garment output image by utilizing the geometric settings in the 3D space.

1 INTRODUCTION

The COVID-19 pandemic impacted 470 million people worldwide by March 2022 in various aspects, including social life and the economy. In the garment industry, in March 2020, Bloomberg reported that approximately 1,089 garment factories in Bangladesh had faced orders worth \$1.5 billion being scrapped (Devnath, 2020). Also, the indefinite closure of factories which leads to unacceptable salary cuts and retrenchments is unavoidable. In response to the pandemic, significant retailers in the United States, such as H&M, Nike, Adidas, etc., greatly reduced their operating hours or even announced their outlet closures. In such scenarios, it is recommended to stay home and do online shopping if possible. This has led to a notable decline in physical shoppers, which in turn increases the number of online shoppers. It can be concluded that the pandemic has remarkably transformed the human lifestyle as well as shopping behavior. For retailers to stand out from other competitors, the key is to provide a favorable environment for e-shopping.

In addition to the pandemic, advanced technology has also made, online shopping possible. According to Statista, internationally, the total amount of goods and services sold online has increased from US\$1.3 trillion in 2014 to US\$2.8 trillion in 2018. Up to 2021, it has further increased to US\$4.9 trillion, and the forecast shows that it will grow by 50% in the coming four years, to approximately US\$7.4 trillion by 2025 (Chevalier, 2022). As such, Statista states that globally as of 2018, the garment industry has generated 57% of the total revenue in e-commerce, with apparel as the most popular category (Chevalier, 2022). Specifically, the VFR plays a vital role in the clothing and garment industry. It not only eases their shopping experience from home with try-on virtually, but it also provides retailers with an opportunity to bridge the gap between online and offline shopping experiences.

Following the trend, the VTO technology embedded in the VFRs has drawn much attention. Numerous research on the technology has been done with uncountable networks introduced, whereby almost all of them are of 2D image-based solutions, that do not resort to any 3D information. Such

^a^{id} <https://orcid.org/0000-0003-2026-5666>

networks include the Virtual Try-On Network (VITON) (Han et al., 2018), UVTON (Kubo et al., 2019), etc., rely only on the Thin Plate Spline (TPS) transformation (Bookstein, 1989) for garment warping, which is known as inaccurate. Undeniably, 2D VTO solutions provide promising results economically as they only involve the reformulation of simple images. Whereas, the 3D VTO solutions such as the Monocular-to-3D Virtual Try-On Network M3D-VTON (Zhao et al., 2021), Clothing Three-Dimensional Reconstruction for Hybrid Image-Based Virtual Try-On (CloTH-VTON+) (Minar & Ahn, 2020), 3D Multiple Pose Virtual Try-On Network 3D-MPVTON (Tuan et al., 2021), are costly to develop as they require high-specification devices for data collection and 3D modeling computations, and/or fancy cameras for physics simulation to capture 3D information underneath, and processing units or sensors for modeling and rendering. Nevertheless, they provide promising fine details in the output with 3D information underneath, enhancing user experiences during garment try-on.

To address the limitations faced, the 3D VFN aims to fit a garment image onto a single-person image, synthetically, with photo-realistic details and deformations well-preserved. This paper introduces a new approach for 3D try-on for garments without utilizing any high-end equipment, that wholly relies on mathematical computation and image processing tasks. With only a garment image and a single-person image as inputs, the network reproduces the warped garment image in 3D with texture, deformations, and any other lifelike information preserved. The 3D VFN is comprised of five stages, the Data Refinement Stage (DRS), the Geometric Matching Stage (GMS), the Depth Estimation and Refinement Stage (DERS), the Try-On Fusion Stage (TFS), and lastly, the 3D Point Cloud Modelling Stage (3D-PCMS). The main contributions of the 3D VFN are:

- Achieved the garment try-on with semantic information well-preserved.
- Achieved the human body reconstruction in 3D space with only a 2D single-person image.
- Improved the geometric matching of try-on quality in terms of alignment and layout adaptation.
- Incorporated the image processing algorithms for detecting the edge within an image based on the image gradients and for alignment and geometric characteristics purposes.

2 LITERATURE REVIEW

2.1 2D VTO Solution

2D VTO solutions solely rely on RGB images with no involvement of 3D information. VITON (Han et al., 2018) first generates a clothing-person agnostic representation with extensive features. The network is equipped with a multi-task encoder-decoder generator to generate a coarse warped garment image with the help of the clothing mask, and a trained network to generate a final warped garment image through composition and refinement processes. To preserve the deformations and comprehensive visual patterns, VITON (Han et al., 2018) implements TPS transformation (Bookstein, 1989) with shape context matching estimation. The standard limitation faced by most VTO solutions is that they are only applicable to images with human models in an upright position. To overcome this, UVTON (Kubo et al., 2019) implemented ultraviolet (UV) mapping for various postures, ensuring high-quality transformation for the geometric information. The texture mapping stream utilizes UV mapping technology with two assistive modules for painting and refinement. It implements DensePose for the estimation and mapping of points of human pixels in 2D RGB images to 3D human module surfaces. The points are mapped to the correlated points of the consumer following the UV coordinate information found in the IUUV, generating a highly defined body part. High Fidelity Virtual Try-On Network via Semantic Adaptation (VTON-HF) (Du et al., 2021) proposed a Semantic Map-based Image Adjustment Network (SMIAN) which aggregates the component features and reconstructs the component images through semantic mapping to generate aggregated body components. To get rid of the texture occlusion and confusion in the semantic mapped result, the component synthesizer interlaces the processed components with that obtained from reference images earlier, to provide a result to be further optimized by the SMIAN loss.

2.2 3D VTO Solution

3D VTO solutions are believed to be more effective because of the 3D information. M3D-VTON (Zhao et al., 2021) reconstructs a 3D try-on mesh by taking only a garment image and a person image as inputs. It proposes a self-adaptive pre-alignment to transform the garment image to be deformed with the TPS transformation, providing a clothing-agnostic person representation and a double-depth map. The map is refined with the help of the shadow information to





- Geometric Matching Stage (GMS) coloured in yellow. This stage is responsible for the garment-person alignment and texture mapping.
- Depth Estimation and Refinement Stage (DERS) coloured in orange. This stage is responsible for human body depth generation and refinement.
- Try-On Fusion Stage (TFS) coloured in blue. This stage is responsible for garment warping to generate a 2D try-on result.
- 3D Point Cloud Modelling Stage (3D-PCMS), coloured in brown. This stage is responsible for the unprojection of RGB-Depth representation of the try-on result to 3D point cloud data and 3D point cloud modeling.

3.1 Data Refinement Stage (DRS)

This stage shown in Figure 1, begins with body posture estimation with OpenPose with the Body 25 model. OpenPose comprises a two-branch multi-stage CNN architecture, which generates the output in two forms: a 2D image pose map and JSON key point coordinates. The pose map is then passed through a feature encoder for 3D human body reconstruction and semantic body parts segmentation tasks. The 3D human body reconstruction adopts the Multi-Level Pixel-Aligned Implicit Function (Saito et al., 2020) that predicts the normal maps for both frontside and backside of the human detected in the image, then reconstructs the human body in 3D space. The semantic body parts segmentation proposes an approach like DeepLabv3+ that shows a huge improvement to DeepLabv2 which produces inaccurate and noisy results. The proposed approach implements the Atrous Spatial Pyramid Pooling (ASPP) scheme and cascades an additional decoder module. The proposed approach adapts the Xception model (Chen et al., 2018) for its proven outstanding performance and rapid computation as shown in Table 1. Apart from that, it also incorporates the Sobel algorithm on the single-person human image for the edge detection process, generating the image gradients Sobel X and Sobel Y, which are the first-order derivatives of the image in the x- and y-directions, respectively as shown in Figure 2. The proposed model is trained on a custom dataset on Google Colab with 50 epochs, 2,500 steps, and a batch size of 2.

Figure 3 shows the proposed model architecture for semantic body parts segmentation. The ASPP first up samples the atrous convoluted features by a factor of 4, while the decoder simultaneously performs a 1x1 convolution for the low-level features to reduce its channels to prevent outweighing of important

Table 1: Segmented Body Maps with DeepLabv2, and with the Enhanced LIP Dataset with the Proposed Approach.

Input Single-Person Image	DeepLabv2 (Chen et al., 2018)		Proposed Approach
	Look into Person (LIP) (Liang et al., 2018)	Active Template Regression (Liang et al., 2015)	Enhanced LIP
			

features, then both the convoluted features are concatenated together. The concatenated features are then passed through a 3x3 convolution to refine the features. Lastly, the convoluted features are up-sampled by a factor of four.



Figure 2: The Image Gradients, Sobel X (Middle) and Sobel Y (Right) Generated with the 3D Human Body Reconstruction by Implementing the Sobel Filter.

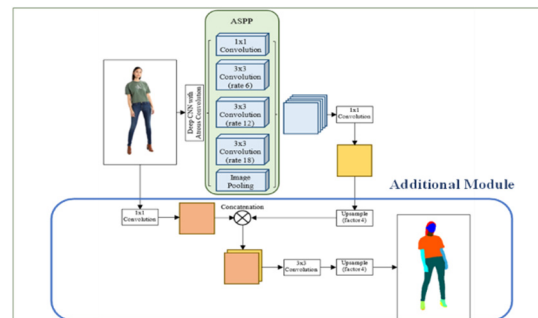


Figure 3: Proposed model architecture for semantic body parts segmentation.

3.2 Geometric Matching Stage (GMS)

Figure 1 first performs the affine transformation to linearly map the garment image to the single-person image for alignment in position- and size-wise purposes:

$$G_{Aff} = \begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix} \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} + \left[\begin{pmatrix} x_{Ap}^M & y_{Ap}^M \end{pmatrix} - \begin{pmatrix} x_{Ag}^M & y_{Ag}^M \end{pmatrix} \right] \quad (1)$$

Whereby G_{Aff} is the affine transformed garment. The relationship indicates the scale factor of the size of the garment image and the person image, such that the garment image is ensured bigger than the person image in size. The midpoints of the pose map and garment image help align the garment image with the pose map to fit the person's size. Subsequently, the aligned affine transformed garment is passed through the TPS transformation network with the person image to obtain a TPS parameter to warp G_{Aff} to the warped garment, transferring the texture and deformations as well. The GMS proposes an improvement on the transfer of geometric details with affine transformation for alignment before TPS transformation, which differs from other VTO solutions. The transformations are presented in Figure 4 and Figure 5.



Figure 4: The Affine Warped Garment, G_{Aff} , with Affine Transformation.



Figure 5: The TPS Warped Garment, G_{TPS} , with TPS Transformation.

3.3 Depth Estimation and Refinement Stage (DERS)

This stage is shown in Figure 1. M3D-VTON (Zhao et al., 2021), introduces an architecture for depth estimation and refinement tasks based on U-Net (Ronneberger et al., 2015). However, it fails to generate fine details, providing a relatively corrupted outcome. As an improvement to M3D-VTON, in this stage, we use the Deep Residual U-Net (Zhang et al., 2018) which reorganizes the U-Net structure with residual connections and an identity mapping path. It comprises three stages: encoder, bridge, and decoder. The stages are made up of residual connections built by two 3×3 convolutional blocks, in which each comprises a convolutional layer, a Batch Normalisation (BN) layer, a Rectified Linear Units (ReLU) activation layer, and an identity mapping path. The encoder first encodes the inputs into several compact representations by applying a stride of two to the first block for feature map halving, instead of feature map down sampling with a pooling operation. With the bridge that connects the encoder to the decoder, the decoder then recovers the compact representations and categorizes them pixel-wise. The feature map is up-sampled within the decoder and concatenated from the encoder before each residual unit. Lastly, the multi-channel feature maps are unprojected with a 1×1 convolutional block and a sigmoid activation layer. The proposed architecture combines the pros of U-Net (Ronneberger et al., 2015) and residual neural network (Zhang et al., 2018), giving a smoothed training process. It also comprises skipped connections between high and low levels of the network, which greatly facilitates information propagation without degradation, reducing the parameters needed. The proposed model is presented in Figure 6. The training of the proposed architecture is done on Google Colab, with 40 epochs, 1,500 steps, and a batch size of 3.

3.4 Try-On Fusion Stage (TFS)

To synthesize a realistic human body texture for 3D human body mesh, this stage implements the Deep Residual U-Net (Zhang et al., 2018) constructed in the previous section to seamlessly merge the warped garment and the single-person image. The synthesis action generates a non-occluded 2D warped garment image with 3D information underneath, that is extracted from the spatial information of the human body along the z-axis embedded in the front depth map. This is shown in Figure 1.

The last stage introduced a try-on fusion network for synthesizing a realistic human body texture. The network architecture proposed for this stage, is similar to that of the DERS, as illustrated in Figure 1. The warped garment and the person image merge and fuse for a seamless fitting. For obtaining the 2D warped garment image, W_{2D} , the synthesis action is guided by the front depth map, segmented body parts map, and the preserved person-part map. On the other hand, the synthesized output also comprises 3D information, which will be unprojected in the next stage. The 3D data is extracted from the spatial information of the human body along with the z-axis lying underneath the front depth map. With the proposed network architecture, a precise try-on result is achieved even for cases with occlusions.

3.5 3D Point Cloud Modelling Stage (3D-PCMS)

The 3D point clouds are obtained by unprojecting the double-depth maps with screened Poisson surface reconstruction (Kazhdan & Hoppe, 2013) as shown in Figure 7. The frontal mesh texture is coloured according to the fitting result, while the back mesh is inpainted with a fast-marching method (Telea, 2004) by filling the backside of the head with a similar hair colour, then mirroring the inpainted image back view to the back mesh. Further image processing procedures are implemented to generate a complete 3D human with a warped garment, which includes computation of surface normals, screened Poisson, and flattening of visible layers.

4 EXPERIMENTAL RESULTS

A pre-trained network from Zhao et al. (Zhao et al., 2021) is adopted. The network is trained on a dataset extracted from randomly extracting images from the For evaluating the performance of the algorithm the 12 test sets are used as shown in Figure 8. The platform used is Python 3.8.13 and PyTorch 1.6.0 for developing, in an Anaconda environment, on PyCharm Integrated Development Environment (IDE).

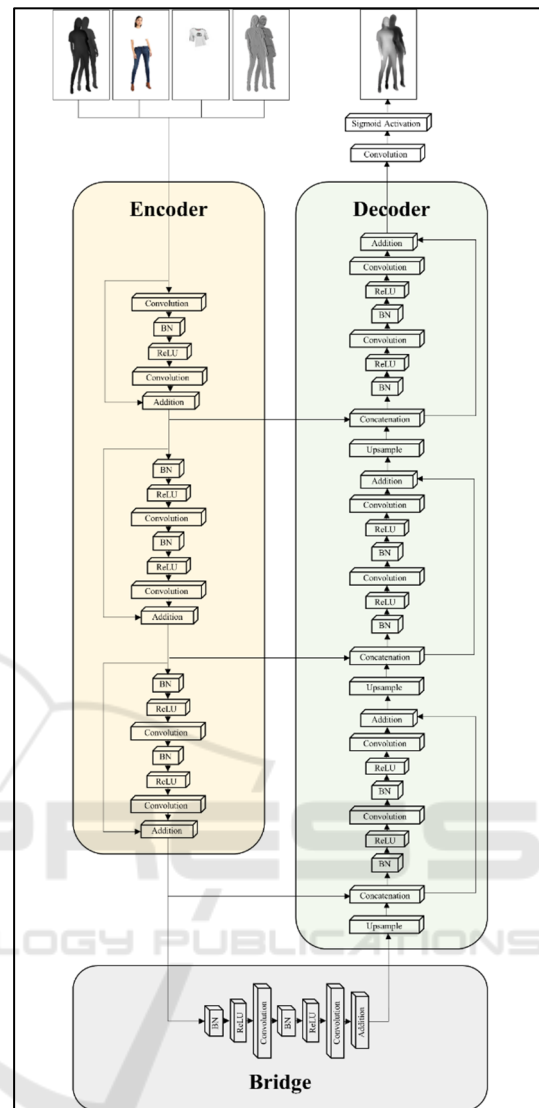


Figure 6: The Deep Residual U-Net Architecture for the DERS.

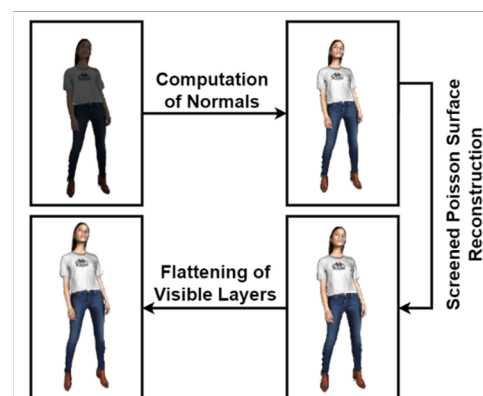


Figure 7: The Image Processing Algorithms Implemented in the 3D-PCMS.



Figure 8: Test Sets (12 Sets, Each Set is Made Up of a Single-Person Image and a Garment Image).

The results obtained show both good and unsatisfying results. For performance evaluation, the results are compared with several methods. As shown in Table 2 and Table 3 by calculating the Structural Similarity Index (SSIM) and Fréchet inception distance (FID). The performance evaluations were performed on both full-body and upper-body try-ons independently.

Table 2: Performance evaluation for full-body try-on.

Network	SSIM	FID
VITON (Han et al., 2018)	0.8861	27.63
UVTON (Kubo et al., 2019)	0.8342	23.11
CloTH-VTON+ (Minar & Ahn, 2020)	0.9012	19.25
3D-MPVTON (Tuan et al., 2021)	0.9134	19.87
3D VFN (Proposed Network)	0.9342	18.42

The quality evaluation for full-body try-on shows that the proposed 3D VFN achieves the highest SSIM of 0.9342, indicating the highest similarity measurement between two images, and the lowest FID of 18.42, indicating the lowest distance measurement between the feature vectors within the image. The quality evaluation for upper-body try-on, on the other hand, shows that the proposed network has the third-highest SSIM and second-lowest FID, which also indicates a satisfying try-on quality with room for improvement.

Table 3: Performance evaluation for upper-body try-on.

Network	SSIM	FID
VITON (Han et al., 2018)	0.8941	27.53
UVTON (Kubo et al., 2019)	0.8843	29.73
CloTH-VTON+ (Minar & Ahn, 2020)	0.8887	27.45
3D-MPVTON (Tuan et al., 2021)	0.8736	27.89
3D VFN (Proposed Network)	0.8857	27.51



Figure 9: Examples of the Unsatisfying Cases of the Proposed 3D VFN.

The proposed network also suffers from several weaknesses as shown in Figure 9. Firstly, the network fails to estimate the spatial gap along the z-direction from 2D RGB images which raises distortions. In addition, the fast-marching inpainting method (TELEA) fails to recognize semantic parts to be inpainted accordingly. The weaknesses are mainly brought by the immaturity of the network trained due to an under-defined dataset. Available datasets only provide garment and single-person images in 2D RGB, without any sideways and spatial information. To achieve higher quality and performance for try-on garments, the dataset for training should be fully furnished with various annotations. Hence, to train the network, the dataset shall comprise garment and single-person images taken from all four directions (front, back, left, and right). Collecting such a large and diverse dataset is challenging as it involves capturing numerous images from different perspectives. Nevertheless, a more mature network can be constructed with such a rich dataset.

5 CONCLUSIONS

This paper introduces a 3D VTO solution, the 3D VFN, which reproduces the warped garment human image in 3D space with photo-realistic information preserved. The proposed network solely relies on 2D RGB images with 24-bit depth and generates a 3D warped garment as output. The network architecture

designed comprises of five stages, which are the Data Refinement Stage (DRS), Geometric Matching Stage (GMS), Depth Estimation and Refinement Stage (DERS), Try-On Fusion Stage (TFS), and 3D Point Cloud Modelling Stage (3D-PCMS), each carries distinct yet significant role. In the architecture, in the DRS, the network first takes in the 2D RGB garment and single-person images as inputs and then refines them into several representations. The GMS performs the affine and TPS transformations for alignment and geometric characteristics transfer purposes. The DERS estimates the human body depth and refines it, followed by the TFS for synthesis action to generate the 2D warped garment human body image. Lastly, the 3D-PCMS models and computes the 3D point cloud of the 3D warped garment human body for finalising it. For assessing the proposed network, SSIM and FID were computed by testing the network on several test sets and the results tabulated show satisfying results and performance.

REFERENCES

- Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6), 567-585.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV),
- Chevalier, S. (2022). Global retail e-commerce sales 2014-2025. *Statista, Key Figures of E-Commerce*.
- Devnath, A. (2020). European retailers scrap \$1.5 billion of Bangladesh orders. *The Guardian*, 23.
- Du, C., Yu, F., Chen, Y., Jiang, M., Wei, X., Peng, T., & Hu, X. (2021). VTON-HF: High Fidelity Virtual Try-on Network via Semantic Adaptation. 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI),
- Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). Viton: An image-based virtual try-on network. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Kazhdan, M., & Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3), 1-13.
- Kubo, S., Iwasawa, Y., Suzuki, M., & Matsuo, Y. (2019). Uvton: Uv mapping to consider the 3d structure of a human in image-based virtual try-on network. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops,
- Liang, X., Gong, K., Shen, X., & Lin, L. (2018). Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4), 871-885.
- Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., & Yan, S. (2015). Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12), 2402-2414.
- Minar, M. R., & Ahn, H. (2020). Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. Proceedings of the Asian conference on computer vision,
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18,
- Saito, S., Simon, T., Saragih, J., & Joo, H. (2020). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
- Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1), 23-34.
- Tuan, T. T., Minar, M. R., Ahn, H., & Wainwright, J. (2021). Multiple pose virtual try-on based on 3d clothing reconstruction. *IEEE Access*, 9, 114367-114380.
- Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749-753.
- Zhao, F., Xie, Z., Kampffmeyer, M., Dong, H., Han, S., Zheng, T., Zhang, T., & Liang, X. (2021). M3d-vton: A monocular-to-3d virtual try-on network. Proceedings of the IEEE/CVF International Conference on Computer Vision,