GlamTry: Advancing Virtual Try-On for High-End Accessories

Ting-Yu Chang
Civil and Environmental Engineering
Stanford University
tingyuc@stanford.edu

Seretsi Khabane Lekena
Stanford Center for Professional Dvelopment
Stanford University
sklekena@stanford.edu

Mothana Alsoofi
Computer Science
Stanford University
mothana@stanford.edu

1. Introduction

The proposed project aims to address the lack of photo-realistic virtual try-on models for accessories such as jewelry and watches, which are particularly relevant for online retail applications. While existing virtual try-on models focus primarily on clothing items, there is a gap in the market for accessories. Leveraging techniques from 2D virtual try-on research, the project will develop a model capable of generating scene-aware try-on images, ensuring realistic and accurate representations of accessories in various contexts. Drawing from relevant literature, the project will customize and retrain the established model with accessory-specific data and modifications to the network architecture.

1.1. Literature Review

Jeongho Kim et al. introduce StableVITON [10], an approach aimed at enhancing the applicability of pre-trained diffusion models for image-based virtual try-on tasks. The primary focus is on preserving clothing details while leveraging the robust generative capabilities of these models. StableVITON achieves this by learning semantic correspondence between clothing and human body within the latent space of the pre-trained diffusion model. Notably, the proposed zero cross-attention blocks maintain clothing details and generate high-fidelity images by incorporating the pre-trained model's inherent knowledge during the warping process. Furthermore, the introduction of a novel attention total variation loss and augmentation techniques enhances the precision of clothing details representation. Through qualitative and quantitative evaluations, StableVITON demonstrates superior performance over baselines, showcasing promising results across diverse images.

Similar to StableVITON, IDM-VTON [7] proposed by Yisol Choi et al. is also a novel diffusion model tailored for image-based virtual try-on tasks. Unlike previous approaches, IDM-VTON effectively preserves garment identity by leveraging two modules to encode garment semantics: 1) high-level semantics integrated into the cross-attention layer, and 2) low-level features incorporated into the self-attention layer. Additionally, detailed textual prompts for both garment and person images enhance the authenticity of the generated visuals. Finally, IDM-VTON presents a customization method using a pair of person-garment images, significantly improving fidelity and authenticity. Experimental results show IDM-VTON outperforming StableVITON and other methods, demonstrating superior fidelity and authenticity in virtual try-on images in a real-world scenario.

In addition, both StableVITON and IDM-VTON utilize VITON-HD [6] as the training dataset to train their respective models. The availability of high-resolution data from VITON-HD is crucial for training robust virtual try-on models capable of synthesizing high-quality images. By leveraging the detailed information provided by VITON-HD, StableVITON and IDM-VTON are able to learn accurate semantic correspondences between clothing items and human bodies, resulting in improved fidelity and authenticity in the generated virtual try-on images. This highlights the importance of high-quality training data like VITON-HD in advancing the state-of-the-art in virtual try-on technology. In this research, we will employ a similar process to create a dataset for accessories. By collecting high-resolution images featuring individuals wearing various accessories such as rings and watches, we aim to compile a diverse dataset that encompasses a wide range of accessory types, poses, lighting conditions, and backgrounds. This dataset will serve as a valuable resource for training and evaluating virtual try-on models tailored specifically for accessories, allowing us to advance the state-of-the-art in this domain.

2. Dataset

In the VITON model family, VITON-HD[6], StableVITON[10], and IDM-VTON[7] all utilize the VITON-HD dataset for training. Seunghwan Choi et al. collected a high-resolution virtual try-on dataset, consisting of 13,679 pairs of frontal-view woman and top clothing images obtained from an online shopping mall website. The dataset was split into a training set with 11,647 pairs and a test set with 2,032 pairs. However, the majority of images in the VITON-HD dataset only feature clothing items, and we were unable to find a publicly available dataset online that sufficiently covers humans wearing accessories. Therefore, this research aims to construct the first comprehensive dataset specifically focused on accessories. We believe that this initiative will greatly benefit future researchers interested in virtual try-on applications for accessories.

2.1. Data Collection

2.1.1 Web Scrapping

To gather new data necessary for training the models towards our objective we built a filtering web scraper using available Google Vision and Search APIs. The first step involves accepting queries to gather images from Google Images and the second step filters them using the google Vision API to remove outliers that do not meet our criteria. Input queries are important here to ensure we gather diverse but relevant images to reduce manual cherry-picking later.

2.1.2 Kaggle

For the try-on items images, we utilized two datasets available on Kaggle. The watch dataset [1] comprises 2000 watch images, categorized into 5 different watch brands. On the other hand, the Tanishq Jewellery Dataset [2] includes images featuring two primary types of accessories: rings and necklaces.

2.2. Data Pre-Processing

After collecting the raw data, we follow the VITON-HD data pre-processing steps. Fig.1 provides an overview of all data types in the VITON-HD dataset. For the StableVITON model, we require agnostic, agnostic-mask, and image-densepose. To obtain these images, we need to complete four steps, including (1) Dense-pose, (2) Human Parsing[12] (3) Pose-estimation[4][14][5][19], (4) Accessories-mask, (5)agnostic-mask and (6) human-agnostic. These steps allow us to construct our three target images.

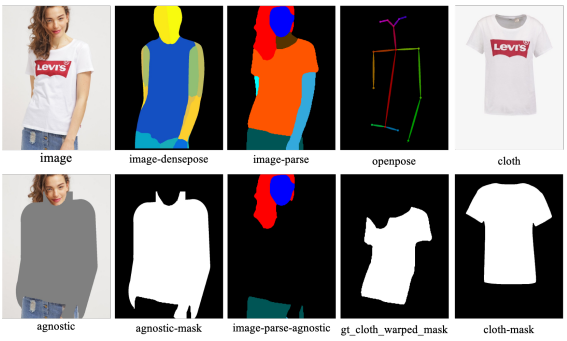

Figure 1: Overview of VITON-HD dataset.

2.2.1 Human Parsing

We experimented with three different models to obtain the Human Parsing feature for the agnostic-mask: (1) Simple Out-of-Box Extractor trained with Look into Person (LIP) dataset [12], (2) Segment Anything (SAM) [11], and (3) SOLOv2 from MMdetection [17]. LIP is the largest single person human parsing dataset with 50000+ images. This dataset focus more on the complicated real scenarios. LIP has 20 labels, including 'Background', 'Hat', 'Hair', 'Glove', 'Sunglasses', 'Upper-clothes', 'Dress', 'Coat', 'Socks', 'Pants', 'Jumpsuits', 'Scarf', 'Skirt', 'Face', 'Left-arm', 'Right-arm', 'Left-leg', 'Right-leg', 'Left-shoe', 'Right-shoe'. In Fig.4 (a), we observed that the Simple Out-of-Box Extractor performs the best in segmenting the watch accurately. However, the watch was classified as background due to the absence of watches in the training labels. SAM also provides accurate segmentation results, but the mask lacks a specific label. In contrast, SOLOv2 categorizes the watch as part of the human.

To address the problem, we plan to retrain the Simple Out-of-Box Extractor with a parsing dataset that includes 12,701 images with manually annotated parsing labels from DeepFashion-MultiModal [9]. The parsing dataset comprises 24 categories, including desirable labels such as ring, wrist wearing, earrings, and necklace. We aim to improve the segmentation results by incorporating this dataset into the training process.

2.2.2 OpenPose

OpenPose [4] pioneered the first real-time multi-person system capable of simultaneously detecting human body, hand, facial, and foot key points in single images, such as {"Nose": 0, "Neck": 1, "RShoulder": 2, "RElbow": 3, "RWrist": 4, "LShoulder": 5, "LElbow": 6, "LWrist": 7, "RHip": 8, "RKnee": 9, "RAnkle": 10, "LHip": 11, "LKnee": 12, "LAnkle": 13, "REye": 14, "LEye": 15, "REar": 16, "LEar": 17, "Background": 18}, totaling 19

Figure 2: The results of human parsing obtained from different models.

key points. Expanding the data points of OpenPose to make use of hand pose estimation did not improve our data set here. Wrist estimations accuracy did not improve against our dataset [4, 14, 5, 19]. We utilize pose estimation features along with human parsing to pre-process human images, transforming them into human-agnostic images for use in the final query data.

As depicted in Fig.4 (b), we observe that the model still predicts key points for the leg or lower body. However, our images primarily focus on the body part above the waist, as we are primarily interested in the upper body or even hand features for accessories. Thus, further data processing may be required, or we can choose to ignore the key points below the waist.

2.2.3 MediaPipe

To enhance the detection of hand features, we incorporate the MediaPipe Hand Landmarker task [20]. This model accurately identifies and localizes 21 key points on the hand (See Fig.3), covering all knuckle coordinates within the detected hand regions. It has been trained on a diverse dataset of approximately 30K real-world images and numerous synthetic hand models superimposed on various backgrounds. Fig.4 (c) shows the results visualization of the hand pose estimation. We select keypoints $0, 9$, and 13 to develop a straightforward algorithm for predicting the watch location. In addition, we save the hand pose estimation as an image to be used as an input for the synthesis generator (See the right image of Fig.4 (c)).

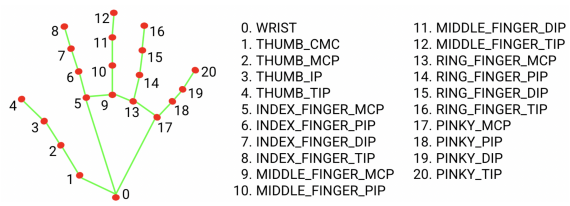

Figure 3: The predicted 21 key points of MediaPipe Hand Landmarker.

2.2.4 Accessories-mask

The accessories mask of the items we are interested in trying on represents the agnostic mask of one of the inputs for the pre-trained U-Net, which serves as the query (Q). We apply U2-net [**?**] to obtain the outline of the items. Fig.4 (d) illustrates the mask result of the target watch for the try-on process.

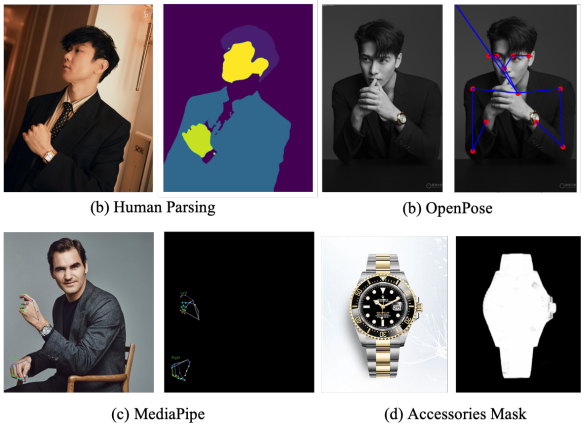

Figure 4: (a) The original image of people wearing watch and its human Parsing output. (b) The original image of people wearing watches and its OpenPose output. (c) The original image of people wearing watches and its MediaPipe output. (d) The original image of the target watch from Kaggle and its mask.

2.2.5 Agnostic-mask and Human-agnostic

Agnostic-mask and Human-agnostic are crucial images for the model's input and can be generated after preparing the previous data types. Our goal is to segment the target accessory mask in human parsing using OpenPose estimation. Specifically, we focus on the labels "RWrist" and "LWrist" of the keypoints stored in the JSON file obtained from OpenPose, as well as the "background" label of the human-parsing mask since the watch is classified under this label.

First, we establish a threshold for the distance between the wrists and create a circular mask with the wrist points as the center. Then, we perform an intersection operation between the background mask and the circle to obtain the desired region. While this approach yields satisfactory results, we observed instances where individuals are positioned close to the camera, resulting in incomplete wrist point predictions, denoted as $(x_{\text{wrist}}, y_{\text{wrist}}) = (0, 0)$. Consequently, the generated mask appears as a circle in the corner (See Fig.5 (a)).

To address this issue, we developed an algorithm that relies solely on the parsing mask to generate the watch mask. Initially, we combine the "Left-arm" and "Right-arm" masks and identify the two largest contours, which should correspond to the same labeled mask separated by the watch. Subsequently, we calculate the bounding boxes (See Fig.5 (c)) of the two contours, along with the coordinates of their centroids. The midpoint between the centroids serves as the estimated location for the missing wrist keypoints. After acquiring the agnostic-mask, we apply it to the original image by overlaying it in gray color. Fig.5 (b) demonstrates a significant improvement in performance achieved with our heuristic algorithm.

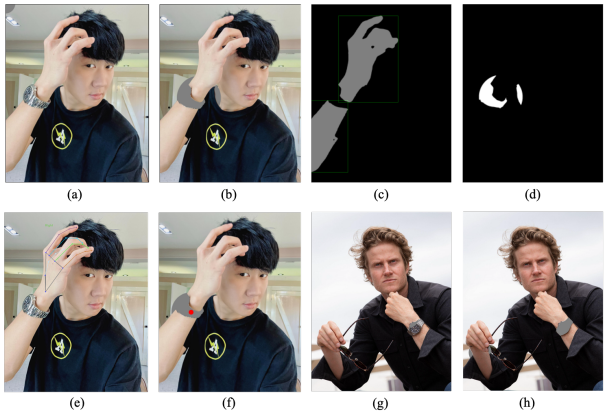

Figure 5: (a) The initial unsuccessful prediction of the human-agnostic mask. (b) The enhanced human-agnostic mask achieved using our algorithm. (c) Arm masks with bounding boxes overlaid. (d) Agnostic mask highlighting the targeted watch region. (e) The original image with predicted MediaPipe Hand Landmarker hand pose. (f) The predicted watch location using the midpoint algorithm. (g) An instance of unsuccessful watch mask prediction. (h) An improved watch mask prediction by the midpoint algorithm.

However, despite the improvement, some of the predicted parsing watch masks remain inaccurate, as shown in Fig.5 (g). To address this, we utilize the MediaPipe Hand Landmarker task, as mentioned in the previous section, in an effort to achieve better results. We begin by finding the midpoint between $Keypoint$ 9 and $Keypoint$ 13, and then treat $Keypoint$ 0 as the midpoint between this midpoint and the watch location. The red dot in the Fig.5 (f) is the predicted watch location. This approach is necessary because the wrist keypoint provided by MediaPipe does not precisely correspond to the actual location of the watch (See Fig.5 (e)). Furthermore, we identified an issue with the heuristic algorithm: it uses a fixed radius for the circle, which is not suitable given the varying image sizes. To address this, we calculate and store the distance between the predicted watch location and $Keypoint$ 0 as the radius for the circle mask. This radius is then used to create the circle, and we perform an intersection with the parse image to refine the mask. Fig.5 (h) illustrates an example of an improved result using the revised algorithm.

3. Method

Our model architecture is primarily based on VITON-HD [6]. An overview of the VITON-HD model is presented in Fig. 6. We skip the segmentation part by utilizing our predicted watch mask. Since we aim to assess the ability to extend the clothing try-on task to watches within the pre-trained VITON-HD framework, we employ the pre-trained Geometric Matching Module model and ALIAS Generator with the data following the pre-processing steps outlined previously as the baseline model to evaluate its performance on watches. Here, we maintain the pose map P as the pose estimation obtained by OpenPose to align with VITON-HD, instead of using the proposed MediaPipe as a baseline.

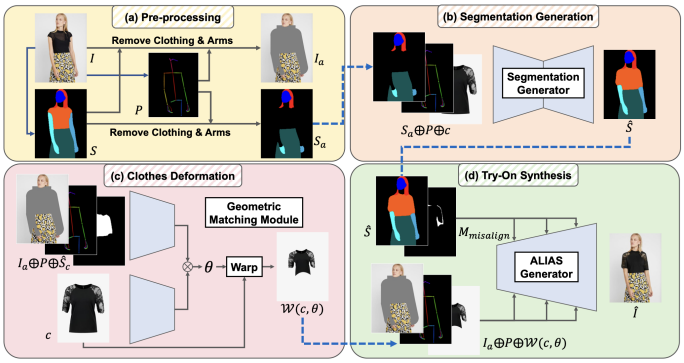

Figure 6: Overview of a VITON-HD. (a) First, given a reference image I containing a target person, we predict the segmentation map S and the pose map P, and utilize them to pre-process I and S as a clothing-agnostic person image I_a and segmentation S_a. (b) Segmentation generator produces the synthetic segmentation $\hat{S}$ from (S_a,P,c). (c) Geometric matching module deforms the clothing image c according to the predicted clothing segmentation $\hat{S}_c$ extracted from $\hat{S}$. (d) Finally, ALIAS generator synthesizes the final output image $\hat{I}$ based on the outputs from the previous stages via our ALIAS normalization.[6]

The baseline results are expected to be poor for two main reasons. Firstly, the entire human pose may not be directly relevant to the position of the watch. Secondly, the model is trained primarily for clothing, meaning that the clothes deformation model may not accurately align the watch with the correct position. Therefore, we need to replace the OpenPose pose map with the output from the MediaPipe Hand Landmarker and subsequently retrain the Geometric

Matching Module (GMM) first proposed in CP-VTON [16] using our watch data. Fig.7 represents our model architecture designed to test our hypothesis for achieving improved results.

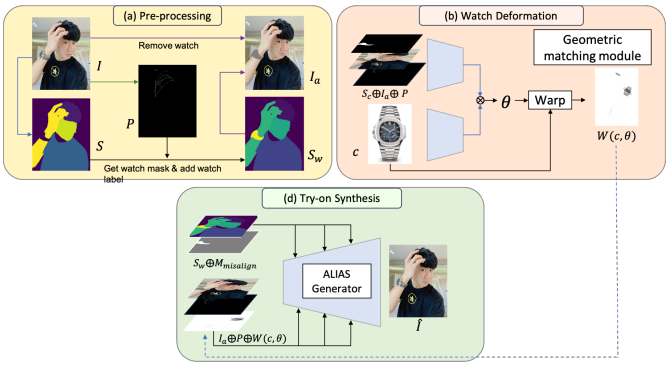

Figure 7

CP-VTON adopted a geometric matching module to learn the parameters of TPS transformation, which improves the accuracy of deformation. Here, we customize CP-VTON+ [**?**] architecture for our task. CP-VTON+ is named after the baseline CP-VTON, outperforms CP-VTON by large margins, in both perceptible and subjective evaluations. The CP-VTON GMM network is built on CNN geometric matching. Whereas the CNN geometric matching uses a pair of color images, CP-VTON GMM inputs are binary mask information, silhouette, and joint heatmap and the colored try-on clothing (See Fig.8). Here we follow VITON-HD to use binary mask of watch, hand pose map and colored try-on clothing (See Fig.9).

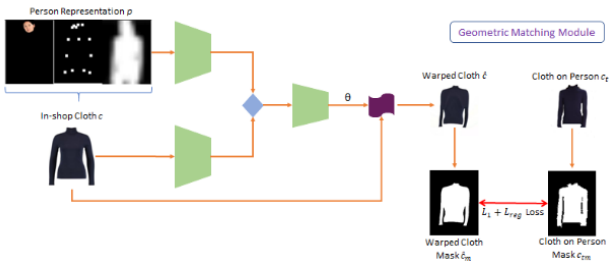

Figure 8: Full GMM pipeline of CP-VTON+.

This GMM (Geometric Matching Module) network, is designed for geometric matching tasks, particularly for aligning two input images spatially. The network consists of several components. First, it employs two feature extraction modules (extractionA and extractionB) to extract high-level features from the input images. These features are then normalized using L2 normalization. The normalized features

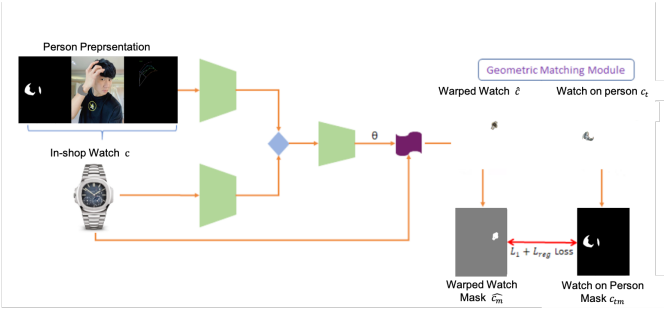

Figure 9: Full GMM pipeline for our task.

are passed through a correlation layer, which calculates the correlation between the features from the two images. This correlation represents the similarity between corresponding spatial locations in the two sets of images. The network then utilizes a regression module to predict the parameters of a thin-plate spline (TPS) transformation based on this correlation. These parameters are used to generate a dense grid of control points. Finally, a TPS grid generator module utilizes these control points to generate a transformation grid, which can be applied to one of the input images to align it with the other image.

As for the loss function, The GicLoss module is used to enforce geometric consistency between neighboring grid points in the generated transformation grid. This loss penalizes differences in distances between neighboring grid points before and after transformation, ensuring smooth and consistent deformation. Overall, the network and loss function work together to learn a transformation that aligns the input images both geometrically and perceptually.

$$\theta = f_\theta\left(f_H(H_t), f_C(Ci)\right) \tag{1}$$

$$L_{GMM}^{CP-VTON+} = \lambda_1 \cdot L_1(C_{\text{warped}}, I_{Ct}) + \lambda_{\text{reg}} \cdot L_{\text{reg}} \tag{2}$$

$$L_{\text{reg}}(G_x, G_y) = \sum_{i=-1,1} \sum_x \sum_y |G_x(x+i, y) - G_x(x, y)| + \sum_{j=-1,1} \sum_x \sum_y |G_x(x, y+j) - G_x(x, y)| \tag{3}$$

In the CP-VTON+ experiments, it was observed that the clothing warping often resulted in significant distortion when compared to existing methods. While the exact cause remained unclear, it was evident that regularization of TPS parameters was necessary to account for the intricacies of clothing textures. To address this, the authors introduced

grid warping regularization, focusing on grid deformation rather than directly manipulating TPS parameters. This approach was chosen for its simplicity and ease of visualization, aiming to minimize abrupt changes in warping between adjacent grid intervals, as depicted in equation 3.

We have decided to utilize the pre-trained ALIAS Generator with 100.5 million parameters in VITON-HD for both the baseline model and our proposed architecture, with a minor modification involving the replacement of the OpenPose map with MediaPipe in the latter. This decision stems from our belief that a well-trained model like ALIAS should be capable of embedding the extracted image features and generating high-quality results based on them, irrespective of variations such as different individuals, clothing items, or watches. One aspect we may question is whether ALIAS has fully grasped the nuanced relationship within P solely through the entire human pose, potentially leading to inaccuracies in location alignment prediction. Fig. 10 shows the details of ALIAS Generator in VITON-HD.

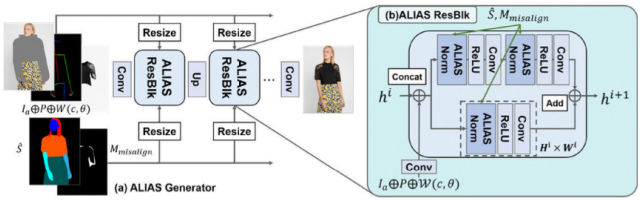

Figure 10: ALIAS generator. (a) The ALIAS generator is composed of a series of ALIAS residual blocks, along with up- sampling layers. The input (I_a, P, $W(c,\theta)$) is resized and injected into each layer of the generator. (b) A detailed view of a ALIAS residual block. Resized (I_a, P, $W(c,\theta)$) is concatenated to h^i after passing through a convolution layer. Each ALIAS normalization layer leverages resized $\hat{S}$ and $M_{misalign}$ to normalize the activation. [6]

4. Experiment

4.1. Baseline Model

Under the architecture of our method as depicted in Fig.7, but utilizing the OpenPose map along with the pre-trained GMM and ALIAS in VITON-HD, we obtain the generated images, warp image, and misalign mask as baseline results. Figure 11 (a), (b), (c), and (d) display the input images for one instance. We create a small test dataset from the images we collected, consisting of paired images of individuals and the watches they are wearing, which the model has not previously seen. These images are resized to 1024×768. This approach benefits the evaluation process. We treat the person images as the ground truth data to compare the differences between these and the generated results.

Using the output of 96.jpg as an example, Figure 11 (e), (f), (g), and (h) illustrate that the warped watch image obtained by the GMM and the output from the ALIAS Generator in VITON-HD preserve good details of both the person and the watch. However, the location prediction does not perform as well as expected, resulting in the agnostic area not properly generating the target watch. Furthermore, we observe that images with non-white or complex backgrounds perform worse than this example. Figure 12 (a) and (b) represent the results of images with complex backgrounds. While we can still discern the outline of the person and other objects, the colors are not preserved, resulting in faded images generated by the generator. We found that there are more instances like this than in the case of 96.jpg. We believe this is because the pre-trained ALIAS model is trained with uniform images featuring a white background; therefore, the model may struggle to effectively handle images with complex backgrounds. However, we are still impressed by the model's ability to preserve most of the image details despite the faded colors.

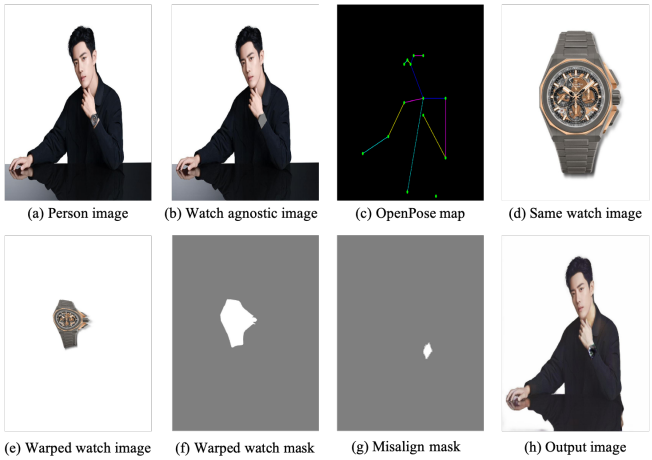

Figure 11: Experiment using the baseline model.

Figure 12: Poor output from the baseline model experiment.

4.2. Model with Watch Deformation

4.2.1 Retraining

It's uncertain whether the issue we're observing in the baseline results stems from incorrect GMM predictions or from the ALIAS model itself. Therefore, we've decided to prioritize retraining the GMM model initially to assess if any improvements can be achieved.

Due to time constraints, we've compiled a small training dataset consisting of hundreds of images for GMM training within the allocated time frame. To monitor the training progress and maintain a record of the training history, we're using TensorBoard. We use the Adam optimizer to update the parameters of our model, with a specified learning rate (lr) set to opt.lr and momentum parameters (betas) of 0.5 and 0.999. Additionally, we implement a step-based learning rate decay strategy, where the learning rate is reduced every 10,000 steps, with the total number of steps before adjustment determined by $keep\_step + decay\_step$.

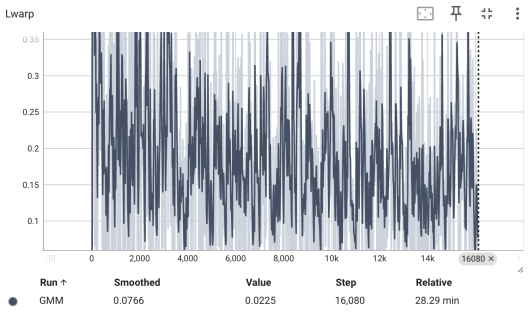

Figure 13: Training history of $\lambda_1 \cdot L_1$

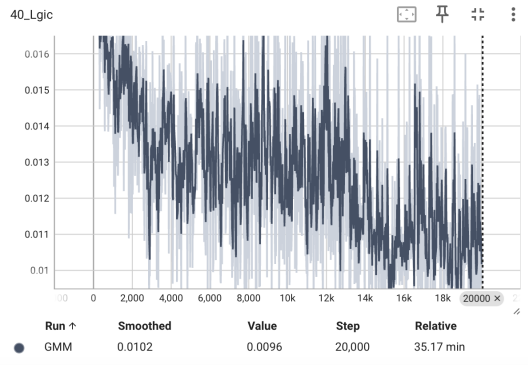

Figure 14: Training history of $\lambda_{reg} \cdot L_{reg}$

Fig.13 and 14 visualize the training history of first and second components of $L_{GMM}^{CP-VTON+}$ in equation 3. It's observed that the loss values for both components oscillate during training. Despite the oscillations, the overall trend of the loss values of $\lambda_{reg} \cdot L_{reg}$ demonstrate a gradual decrease over epochs. However, the loss values corresponding to $\lambda_1 \cdot L_1$ do not exhibit this desirable trend, suggesting the need for potential adjustments in the learning rate or fine-tuning of the weighted coefficient λ_1.

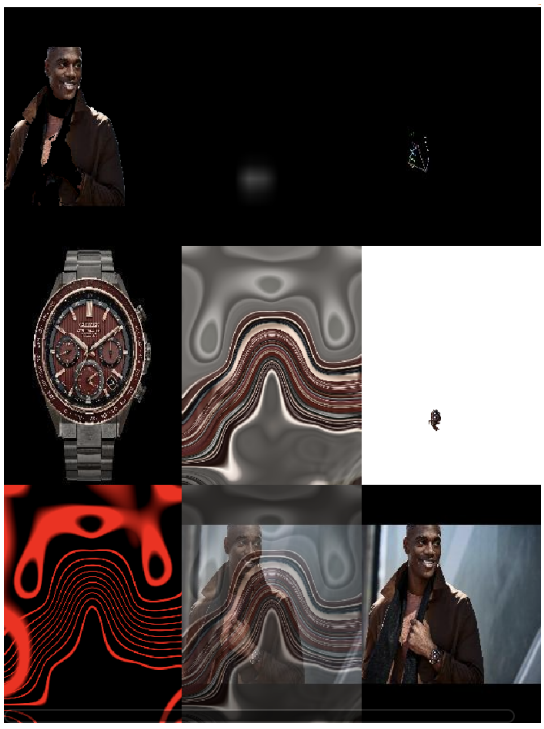

Figure 15: The severely distorted warped grid and image.

We've also encountered a challenge similar to that reported in previous studies, where some of the warped watches are severely distorted. As depicted in Fig.15, a instance showcases this issue, where it's difficult to discern the watch in the center image, and the bottom-left warped image is notably distorted. The color of the background could be a contributing factor to this issue. We noticed that most of the watch images with a dark background exhibit this problem, whereas we haven't encountered it with images featuring a white background.

5. Results/Evaluation

5.1. Results

First, we compare the quality of warped watch generation between the pre-trained GMM in VITON-HD and our re-trained GMM. We observed that our model preserves the shape of the watch better (See Fig.16) and the size is more reasonable relative to the scale of the person in the image. Although the predicted location is still not accurate, we can

see that it is closer to the target area compared to the baseline results. Taking 96.jpg as an example again, Fig.17 illustrates the improvement in trend. We can anticipate this outcome given that we only have a small dataset for training compared to the GMM in VITON-HD, which is trained with approximately 16,000 images. Our dataset is insufficient for the model to learn the relative location relationship between the watch mask of the target watch and the watch-agnostic areas in the person image. This experiment has demonstrated improvement, suggesting further potential enhancement with a larger dataset.

(a) Warped watch for 131.jpg generated by pre-train GMM in VITON-HD and our re-train GMM

(b) Warped watch for 127.jpg generated by pre-train GMM in VITON-HD and our re-train GMM

Figure 16: Comparison of the warped watches generated by pre-train GMM and our re-train GMM.

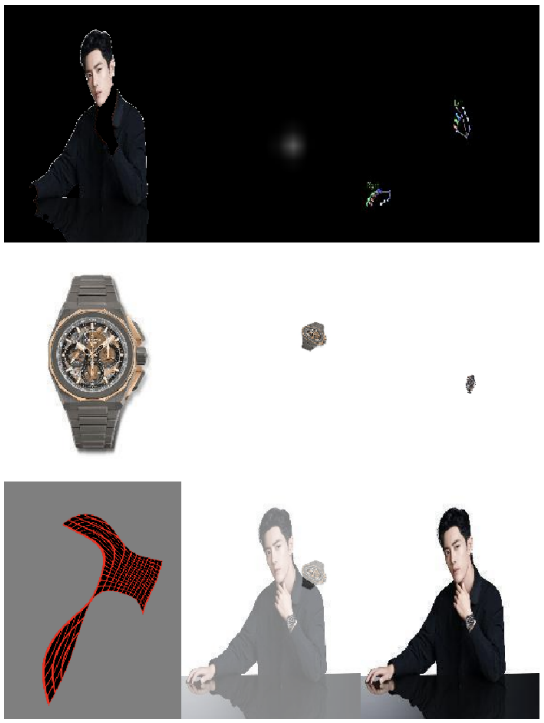

Figure 17

The faded color issue was not resolved by retraining the GMM. Figure 18 compares the baseline results with the proposed method, using the same examples as before. There is little difference between the baseline and the proposed method because the locating prediction is still inaccurate and the ALIAS Generator has not been retrained with the images with complex background.

(a) Baseline results and our results for 120.jpg.

(b) Baseline results and our results for 130.jpg.

Figure 18: Comparison between the output of baseline model and ours.

5.2. Qualitative Results

Since Virtual Try-On is a practical application of generative models, qualitative evaluation can be conducted directly by human observation. Performance can be assessed by comparing images generated by different models, figure 19. Furthermore, we aim to create a simple questionnaire to gather feedback from our friends and classmates on the quality of the generated images.

5.3. Quantitative Results

For quantitative evaluation, we employ SSIM [18] and LPIPS [21] metrics in the paired setting. In the unpaired setting, realism is assessed using FID [8] and KID [3] scores. Our implementation follows the evaluation paradigm [13]. We evaluated 48 images, figure **??** shows their resulting SSIM scores over different resolution sizes. Large was the training size. While the charts look very similar, ours improved the averaged SSIM score. The LPIPS scores were close between our results and the baseline see figure 20.

6. Future Work

Glamtry can be improved by training its dependent models more extensively on our dataset. The reliance on clear head-to-wrist visibility, resulting from the use of OpenPose, causes our model to struggle with replacing watches in common close-up wrist photos. Training the model to accurately reproduce specular details is another avenue forward. This would improve the model for correct jewelry representation during try-on scenarios.

7. Contributions

Ting-Yu Chang, tested various data pre-processing pipelines and worked improving the model's warp components and re-architected the model to no longer include the

segmentation Generation step. Ting-Yu, focus on refactoring the VITON-HD model to generate our baseline and retrain GMM to obatin final results.

S. Khabane Lekena worked researching methods to improve segmentation through hot swapping other segmentation models like PGN. Khabane also wrote the image scraping code. He also implemented existing code for the evaluation steps mentioned. PGN: Instance-level Human Parsing ECCV 2018 Paper

Mothana Alsoofi, explored techniques to enhance model's warp and additional segmentation models with adjusted parameters. Mothana experimented cross-validation techniques on a large test dataset to try to gauge accurate performance assessment by trying other score calculations (like MS-SSIM) for evaluation.

7.1. acknowledgements

We used Pranjal Datta's SSIM notebook [15]

8. Appendix

Figure 19: Qualitative side by side. While the images show strong distortion, ours (right two) shows less distortion on the rest of the body.

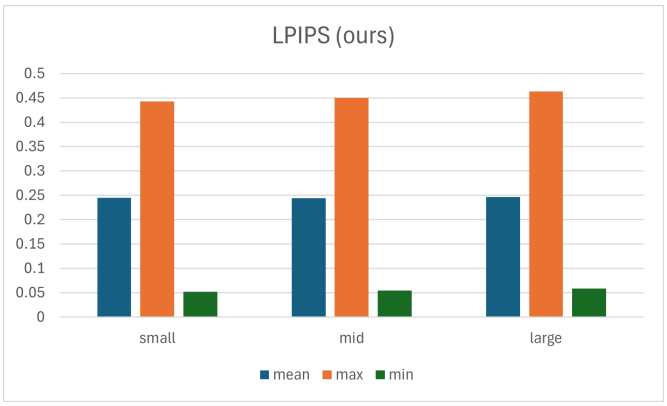

Figure 20: LPIPS scores for our model. The score closely resembles the baseline score. Ours does not match the baseline by a small margin.

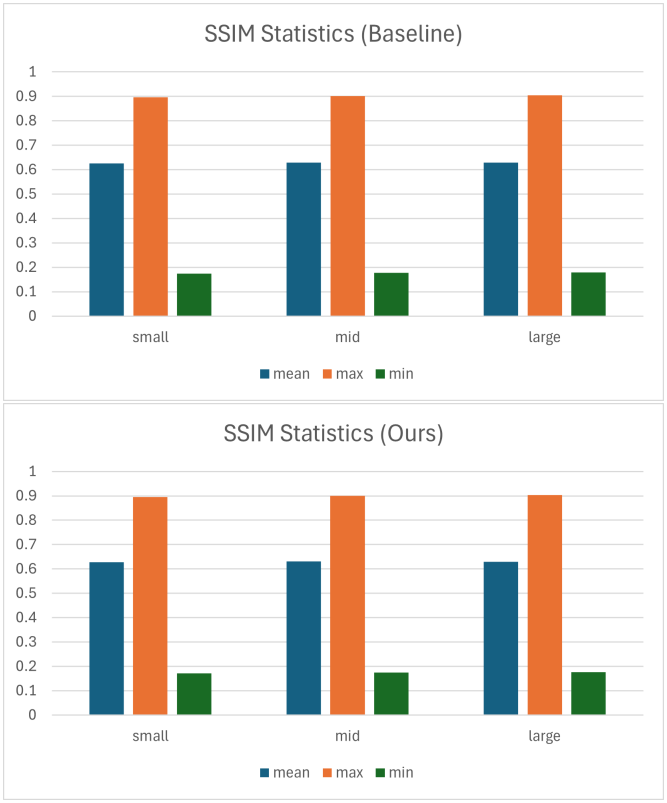

Figure 21: Side by side of the SSIM scores between the baseline (left) and ours (right). Ours had a high average score while the baseline maintained higher min/max scores.

References

[1] Fancy watch images dataset. https://www.kaggle.com/datasets/ahedjneed/fancy-watche-images. Accessed: 2024-06-08. 2

[2] Tanishq jewellery dataset. https://www.kaggle.com/datasets/sapnilpatel/tanishq-jewellery-dataset. Accessed: 2024-06-08. 2

[3] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 8

[4] Z. Cao, G. H. 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. 2, 3

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 3

[6] S. Choi, S. Park, M. Lee, and J. Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021. 1, 2, 4, 6

[7] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 1, 2

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8

[9] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2

[10] J. Kim, G. Gu, M. Park, S. Park, and J. Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *arXiv preprint arXiv:2312.01725*, 2023. 1, 2

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[12] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[13] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara. Ladi-vton: latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023. 8

[14] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2, 3

[15] SRM-MIC. All about structural similarity index (ssim) - theory & code in pytorch. https://medium.com/srm-mic/all-about-structural-similarity-index-ssim-theory-code-in-pytorch-6551b455541e, 2023. Accessed: 2024-06-08. 9

[16] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 5

[17] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen. Solov2: Dynamic and fast instance segmentation. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2, 3

[20] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann. Mediapipe hands: On-device real-time hand tracking, 2020. 3

[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8