Literature Review for Virtual Try-On

Aditya Bhat (ab2260), Anirudh Negi (an721), Harsh Patel (hkp49)

February 28, 2022

There are two main categories for virtual try-on approaches: 3D model-based approaches and 2D image-based approaches. 3D model-based approaches can accurately simulate the clothes but are not widely applicable due to their dependency on 3D measurement data.

Image-based Virtual Try-On (VITON)[1] is a 2D based Network, which uses coarse-to-fine strategy by generating synthesized image and augmenting the cloth on to the person maintaining the pose. It involves has 3 parts: Person Representation where pose heatmap, human body representation and face and hair segment are used. Next, It transfers the target fashion item depending on the location of different body parts (e.g., arms or torso) and the body shape. Finally, to maintain the identity of the person, they incorporate physical attributes like face, skin color,hair style, etc. One of the advantage of this model is that it uses body shape information along with target size to handle occlusion and pose ambiguity resulting in more comprehensive and effective clothing-agnostic representation. VITON also accurately generates detailed virtual try-on results. However, it cannot determine which regions near the neck should be visible and also fails to rarely-seen poses or a huge mismatch in the current and target clothing shapes. One of the other limitation is the imperfect shape-context matching for aligning clothes and body shape, and the inferior appearance merging strategy.

VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization[2] This is the paper we have decided to follow in our project. Approach involves mainly 4 stages. Pre-processing: In this stage we try to eliminate torso and arm (excluding the wrist). Segmentation Generation: Segmentation generator is a Neural network which gives us a segmentation map as an output in which the person is wearing the cloth. Clothes deformation: In this step, we deform our reference cloth as per the personâs pose. Try-on synthesis: In our final step we try to combine all, i.e output of the segmentation generator and and wrapped cloth along with output of first step, is fed into a ALIAS generator (Alignment Aware Segment Normalization). This enables the preservation of semantic information, and the removal of misleading information from the misaligned regions. This model clearly preserves the details of the target clothes, such as the logos and the clothing textures, due to the multi-scale refinement at a feature level. In addition, regardless of what clothes the person is wearing in the reference image, our model synthesizes the body shape naturally

CP-VTON[3] focuses on transforming the target clothes into the most fitting shape seamlessly but also preserve well the clothes identity in the generated image, that is, the key characteristics (e.g. texture, logo, embroidery) that depict the original clothes. The method involves following steps: First, it learns a thin-plate spline transformation for transforming the in-shop clothes into fitting the body shape of the target person via a new Geometric Matching Module (GMM) rather than computing correspondences of interest points as prior works did. Second, to alleviate boundary artifacts of warped clothes and make the results more realistic, and employ a Try-On Module that learns a composition mask to integrate the warped clothes and the rendered image to ensure smoothness. Even though the warped clothes are roughly aligned with target person, CP-VTON(w/o mask) still loses characteristic details and produces blurry results. This verifies that encoder-decoder network architecture like UNet fails to handle even minor spatial deformation. This methodology fails at (1) improperly preserved shape information of old clothes, (2) rare poses and (3) inner side of the clothes indistinguishable from the outer side.

ACGPN[4] is a visual try-on network model which generates photo-realistic try-on and rich clothing details in the resulting image. It involves three major steps. First, a semantic layout generation module utilizes semantic segmentation of the reference image to progressively predict the desired semantic layout after try-on. Second, a clothes warping module warps clothing images according to the generated semantic layout, where a second-order difference constraint is introduced to stabilize the warping process during training. Third, an inpainting module for content fusion integrates all information (e.g. reference image, semantic layout, warped clothes) to adaptively produce each semantic part of human body. In order to maintain the person's shape ACGPN employs the detailed body shape mask as the input, and the neural network attempts to discard the clothing information to be replaced. However, since the body shape mask includes the shape of the clothing item, neither the coarse body shape mask nor the neural network could perfectly eliminate the clothing information. As a result, the original clothing item that is not completely removed causes problems in the test phase.

MG-VTON[5] generates a new person image after fitting the desired clothes into the input image manipulating human poses. It has three stages: a target image is synthesized for desired person to match both the pose and the target clothes shape; a deep Warping Generative Adversarial Network (Warp-GAN) warps the desired clothes appearance into the synthesized human parsing map and alleviates the misalignment problem between the input and the desired human pose; a refinement render utilizing multi-pose composition masks recovers the texture details of clothes and removes some artifacts. The model synthesizes a photo-realistic image by transferring the desired clothes onto the person, which is computationally efficient. It is claimed to outperform all state-of-the-art methods both qualitatively and quantitatively with promising multipose virtual try-on performances. However, it fails to build up the photo-realistic virtual try-on system for the real-world scenario, partially ascribing to the semantic and geometric differences between the target clothes and reference images, as well as the interaction occlusions between the torso and limbs. Also, even though it preserves the characteristics of the clothes, such as texture, logo, embroidery, yet the content generation and preservation remain an uninvestigated problem in this model.

VTNFP[6] is another Image-based virtual try-on system which aims to preserve clothing and body features. VTNFP consists of three modules: a clothing deformation module which transforms cloth to a warped version that aligns with the posture of the person; a segmentation map generation module which generates a new segmentation of body parts as well as body regions covered by the target clothing; and a try-on synthesis module which synthesizes the final target image. The body segmentation map prediction module provides critical information to guide image synthesis in regions where body parts and clothing intersects. VTNFP introduces several methodological innovations to improve the quality of image synthesis, and demonstrates that the method is able to generate substantially better realistic looking virtual try-on images than the state-of-the-art methods. Although it uses segmentation representation to preserve the non-target details of body parts and bottom clothes, it is still inadequate to fully preserve the details, resulting in blurry output. This is due to unawareness of the semantic layout and relationship within the layout, therefore being unable to extract the specific region to preserve. It also cannot avoid distortions and misalignments on the Logo or embroidery, remaining a large gap to photo-realistic try-on.

Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[7] helps in understanding how to render a content image in the style of another image. The deep neural networks encode not only the content but also the style information of an image. The paper particularly explains neural style transfer algorithm that provides the flexibility of the optimization-based framework and the speed similar to the fastest feed-forward approaches. The major advantage is the network tries to synthesize the most stylized image when , a smooth transition between content-similarity and style-similarity can be observed. The major disadvantage is this method is not optimized for texture based high quality images and requires advance neural architectures.

The Conditional Analogy GAN Swapping Fashion Articles on People Images[8] proposes a novel method to solve image analogy problems: Conditional Analogy Generative Adversarial Network (CAGAN) is based on adversarial training and uses deep convolutional neural networks It is applied to automatic swapping of clothing on fashion model photos. The major advantage is this method uses relatively large dataset with 15000 poses to train their model. This model of GAN can swap clothes on humans and offers new possibilities to image manipulation for fashion purposes. The major disadvantage is this method cannot capture the poses completely and to fill in the gaps we need to recreate those areas and color it with the swapped clothes.

ClothFlow[9] is an appearance flow based generative model which synthesize person using its pose for image generation and virtual try-on. It estimates a dense flow between source and target clothing regions and effectively model the geometric changes and naturally transfers the appearance to synthesize novel images. The processing involves three stages: First, it estimates a person semantic layout to provide richer guidance to the generation process. Next is the clothing flow estimation stage, which predicts the appearance flow. Finally, a cloth preserving rendering stage synthesizes the target image with a generative network while trying to preserve details from the warped source clothing regions. The estimated flow properly handles the geometric deformation as well as occlusions/invisibility between the source and target image, making Clothflow favorable to other state-of-the-art methods on two standard image synthesizing tasks. However, Clothflow has the following limitations: It relies on a third party human parser and fails to generate realistic synthesized results when the parsing results are inaccurate. Secondly, without adversarial training, many non-clothing regions do not look realistic such as faces, shoes, etc.

Fashion++[10] proposes a model which requires minimal edits to a full-body clothing outfit to have maximum impact on its fashionability. It uses a deep image generation neural network that learns to synthesize clothing conditioned on learned per-garment encodings which are explicitly factorized according to shape and texture, thereby allowing direct edits for both fit and patterns respectively. Their method involves image generation Framework, which decomposes outfit images into their garment regions and factors shape/fit and texture. Next is the editing module for revising an input's features to improve fashionability. Finally, it uses an activation maximization-based outfit editing procedure to show how the model recommends garments. Their training is scalable in terms of supervision and adaptability to changing trends. Also, the model captures subtle visual differences and the complex synergy between garments that affects fashionability. However, a minimal edit requires good outfit generation models, an accurate fashionability classifier, and robust editing operations. Failure in any of these aspects can result in worse outfit changes.

TailorNet[11] adapts to a user's style by predicting clothing deformation in 3D using pose, shape, and style preserving wrinkle details. Their method involves four steps: Un-posing Garment Deformation for decomposition of clothing as non-rigid and articulated deformation, subspace of garment styles which generates variation in a pose, single style shape model to predict reasonable garment fit along with fine-scale wrinkles and generalizes well to unseen poses, and finally decomposing the garment mesh vertices, in an unposed space into a smooth low-frequency shape and a high frequency shape with diffusion flow. It tries to address the limitations of existing methods which shows overly smooth and non-realistic images by predicting low and high frequencies separately. It also produce temporally coherent deformations as well despite being trained on static data poses from a different dataset. However, the fixed number of shape-style specific mixtures makes the method sensitive towards the number of such components. In addition, since every one of the few shape-style explicit models advanced independently, it invalidates the point of together displaying every one of the varieties in style, body shape, and stances.

Large Scale GAN Training for High Fidelity Natural Images[12] Synthesis helps to generate more samples of the images with minor tweaks. This helps in fetching various poses for virtual trial and creating a dataset that can be used to build a better model for swapping clothes in a virtual try-on environment. The major short-coming is training procedure for convolutional neural networks is brittle without finely-tuned hyperparameters and architectural choices.

References

[1] Xintong Han et al. "VITON: An Image-based Virtual Try-on Network". In: *CoRR* abs/1711.08447 (2017). arXiv: 1711.08447. URL: http://arxiv.org/abs/1711.08447.

[2] Seunghwan Choi et al. "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization". In: *CoRR* abs/2103.16874 (2021). arXiv: 2103.16874. URL: https://arxiv.org/abs/2103.16874.

[3] Bochao Wang et al. "Toward Characteristic-Preserving Image-based Virtual Try-On Network". In: *CoRR* abs/1807.07688 (2018). arXiv: 1807.07688. URL: http://arxiv.org/abs/1807.07688.

[4] Han Yang et al. "Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content". In: *CoRR* abs/2003.05863 (2020). arXiv: 2003.05863. URL: https://arxiv.org/abs/2003.05863.

[5] Haoye Dong et al. "Towards Multi-pose Guided Virtual Try-on Network". In: *CoRR* abs/1902.11026 (2019). arXiv: 1902.11026. URL: http://arxiv.org/abs/1902.11026.

[6] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. "VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation". In: (2019), pp. 10510–10519. DOI: 10.1109/ICCV.2019.01061.

[7] Xun Huang and Serge J. Belongie. "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization". In: *CoRR* abs/1703.06868 (2017). arXiv: 1703.06868. URL: http://arxiv.org/abs/1703.06868.

[8] Nikolay Jetchev and Urs Bergmann. "The Conditional Analogy GAN: Swapping Fashion Articles on People Images". In: (2017). arXiv: 1709.04695 [stat.ML].

[9] Xintong Han et al. "ClothFlow: A Flow-Based Model for Clothed Person Generation". In: (2019), pp. 10470–10479. DOI: 10.1109/ICCV.2019.01057.

[10] Wei-Lin Hsiao et al. "Fashion++: Minimal Edits for Outfit Improvement". In: *CoRR* abs/1904.09261 (2019). arXiv: 1904.09261. URL: http://arxiv.org/abs/1904.09261.

[11] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. "TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style". In: *CoRR* abs/2003.04583 (2020). arXiv: 2003.04583. URL: https://arxiv.org/abs/2003.04583.

[12] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *CoRR* abs/1809.11096 (2018). arXiv: 1809.11096. URL: http://arxiv.org/abs/1809.11096.