Virtual Try-On Review and Summary

[redacted]박윤준[1]

I. ABSTRACT

A. Interested problem

My research interest is in image-based virtual try-on. Virtual try-on systems have revolutionized the online shopping experience by enabling users to visualize how clothes might look on them without physically trying them on.

B. Approach

In this paper, I will explore the challenges faced by image-based virtual try-on tasks, the various approaches employed to address these issues, and how these approaches have subsequently enhanced the performance of virtual try-on tasks.

C. key results

The adoption of Virtual Try-On (VTO) services is increasing among renowned fashion, accessories, and beauty companies to boost e-commerce. These user-friendly services overcome the limitations of not being able to physically try on products online, allowing users to virtually fit clothing or apply makeup more easily than in offline settings. While academia is actively developing models for these technologies, the fashion and beauty industries are already deploying these services extensively.

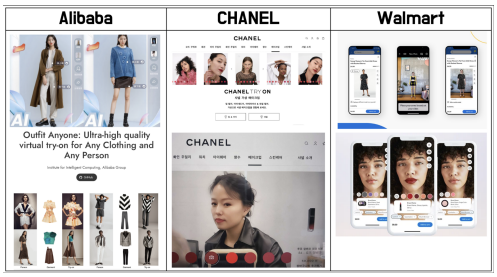

Fig. 1. Prominent fashion, accessories, and beauty companies that are adopting Virtual Try-On services

II. INTRODUCTION

A. Interested problem

The ability to virtually try on clothes has become increasingly important with the rise of e-commerce. Virtual try-on systems use image synthesis techniques to superimpose clothing items onto images of users. This technology not only enhances user experience but also has significant commercial potential. Recent advancements have focused on improving the realism and accuracy of these virtual try-on systems, addressing challenges such as pose variation, garment fit, and image quality.

B. Similarities and Differences in Approaches

The papers I have chosen are: 1. "Image Based Virtual Try-on Network from Unpaired Data," 2. "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization," and 3. "TryOnDiffusion: A Tale of Two UNets."

First, the similarities are: "Image-based virtual try-on aims to synthesize a naturally dressed person image with a clothing image. To be more specific, the task of image-based virtual try-on aims to transfer a target clothing item onto the corresponding region of a person, which is commonly tackled by fitting the item to the desired body part and fusing the warped item(합성하려는 아이템) with the person."

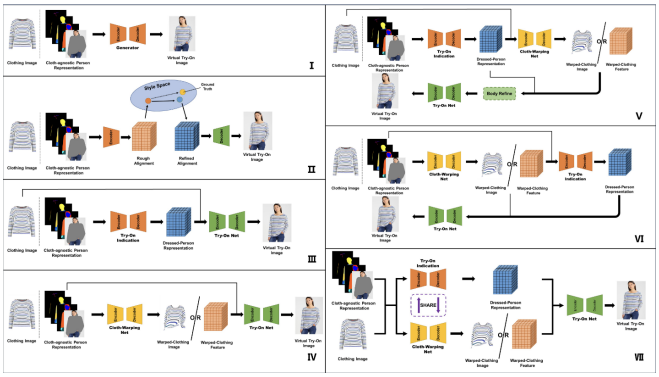

Fig. 2. Basic pipelines(approaches) of image-based virtual try-on

Second, the differences are: "VITON-HD focuses on high-resolution image synthesis, the paper 'Image Based Virtual Try-On Network from Unpaired Data' presents an innovative approach to virtual try-on using image data without the need for paired images, and the paper 'TryOnDiffusion: A Tale of Two UNets' proposes a novel approach to virtual apparel try-on, aiming to address the significant challenges in synthesizing photorealistic visualizations of garments on different body shapes and poses."

C. Approaches to solving the problems

The selected papers vary in their approach to solving virtual try-on problems.

1) Image Based Virtual Try-on Network from Unpaired Data :

a) Shape Generation Network: This network creates a new segmentation map by combining the body shape of the query image (the person) with the shape features of the reference garments. This is done using an autoencoder that generates shape feature slices for different garment types.

b) Appearance Generation Network: This network generates a photo-realistic image by combining the appearance features of the selected garments with the shape map produced in the previous step. The appearance features are extracted using an autoencoder from reference images, and then these features are mapped to the segmentation regions.

c) Online Optimization: This step refines the generated image to improve the quality of fine details such as textures and logos.

d) Why These Approaches are Appropriate :

– Dataset Efficiency

The approach only requires single images of garments and people, significantly reducing the data collection burden compared to methods that require paired images or 3D data.

– Composability

The ability to compose multiple garments into a single outfit makes the method highly flexible and user-friendly, addressing the common scenario of wanting to see an entire outfit rather than a single garment.

– Fine Detail Accuracy

The online optimization step ensures that fine details are accurately rendered, improving the realism and user experience.

e) Datasets Used : dataset of people(both males and females) in various outfits and poses, that we scrapped from the Amazon catalog. The dataset is partitioned into a training set and a test set of 45K and 7K images respectively. All the images were resized to a fixed 512 × 256 pixels.

Fig. 3. Image Based Virtual Try-On Network From Unpaired Data O-VITON algorithm

2) VITON-HD:

- 3D Model-Based Approaches: These methods accurately simulate clothing on 3D models but require detailed 3D measurements, making them less practical for widespread use.
- 2D Image-Based Approaches: These are computationally efficient and suitable for practical use as they do not rely on 3D data. VITON-HD falls into this category, building on previous work such as CAGAN, VITON, and CP-VTON, but with improvements in resolution and accuracy.

a) Key Components of VITON-HD:

– Clothing-Agnostic Person Representation

This involves generating a person image without the clothing item to be replaced, ensuring the preservation of body shape and pose information while removing the original clothing.

– Segmentation Generation

The model predicts the segmentation map of the person wearing the target clothing.

– Clothing Image Deformation

The target clothing is deformed to align with the segmentation map.

– ALIAS Normalization and Generator

The ALIAS normalization addresses misalignment issues by computing separate statistics for misaligned and non-misaligned areas, preserving clothing details and improving image quality.

b) Why These Approaches are Appropriate(Effective) : These approaches are effective because they tackle the core challenges of virtual try-on systems: removing original clothing accurately, handling misalignment, and producing high-resolution, photo-realistic images. The methods leverage deep learning techniques, such as conditional generative adversarial networks (cGANs) and novel normalization layers, to achieve these goals.

c) Datasets Used : VITON-HD uses a newly collected dataset of 1024×768 resolution images consisting of pairs of a person and a clothing item.

- 1024×768의 Virtual Try-on 데이터셋을 수집하였다.(기존 데이터셋은 해상도가 낮다)
- 사람과 의류 이미지의 쌍을 사용하여 paired setting을 평가하고, unpaired setting을 위해 의류 이미지를 섞어 평가한다.
- Paired setting : 사람 이미지를 원래의 의류 아이템으로 재구성하는 것이다.
- Unpaired setting : 사람 이미지의 의류 아이템을 다른 아이템으로 변경하는 것이다.

d) Key Ideas and Skills : Key ideas include the introduction of clothing-agnostic person representation, the use of segmentation maps for guiding synthesis, and the development of ALIAS normalization to address misalignment. Skills involved are deep learning, image processing, and generative modeling.

3) TryOnDiffusion: A Tale of Two UNets: The TryOnDiffusion approach leverages the following key elements:

a) Parallel-UNet Architecture : Combines garment warping and person blending into a single process using cross-attention mechanisms, improving the preservation of garment details even with significant pose and shape variations.

b) Diffusion Models : Utilizes cascaded diffusion models for generating high-resolution images (1024x1024 pixels). This includes a base diffusion model for 128x128 res-

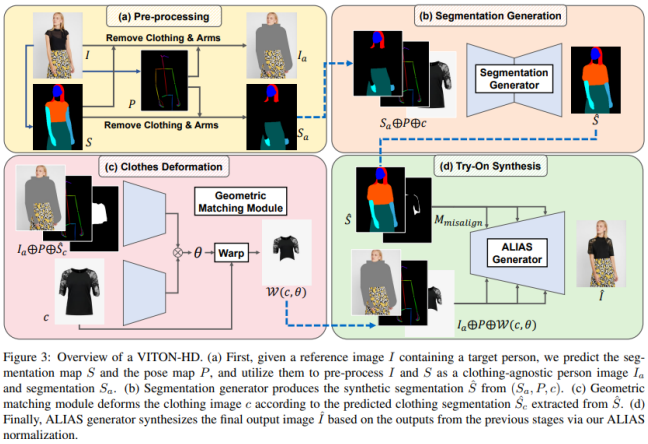

Figure 3: Overview of a VITON-HD. (a) First, given a reference image I containing a target person, we predict the segmentation map S and the pose map P, and utilize them to pre-process I and S as a clothing-agnostic person image I_a and segmentation S_a. (b) Segmentation generator produces the synthetic 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.

Fig. 4. Overview of a VITON-HD

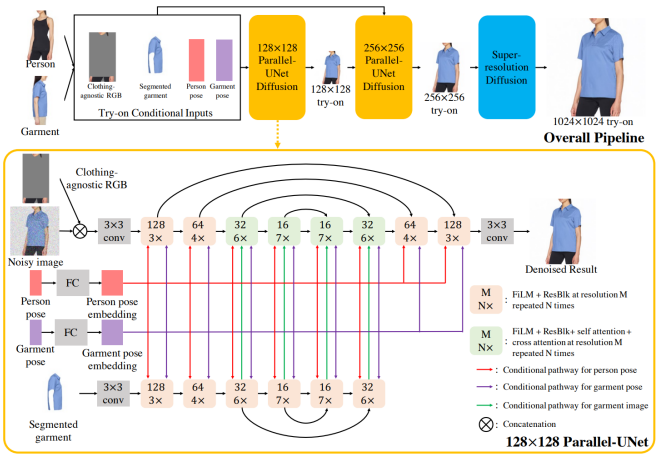

Fig. 5. Overall pipeline

olution, a super-resolution model for 256x256, and another for 1024x1024.

c) Large-scale Dataset : Trains the model on 4 million image pairs, enhancing its capability to generalize across diverse body poses and garment types.

d) Why These Approaches are Appropriate(Effective) : These approaches are effective because they address the common issues in virtual try-on systems, such as handling non-rigid deformations and occlusions, while maintaining high-resolution garment details.

e) Datasets Used: The paper collects a paired training dataset of 4 Million samples. Each sample consists of two images of the same person wearing the same garment in two different poses. For test, we collect 6K unpaired samples that are never seen during training. Each test sample includes two images of different people wearing different garments under different poses. Both training and test images are cropped and resized to 1024×1024 based on detected 2D human poses. Our dataset includes both men and women captured in different poses, with different body shapes, skin tones, and wearing a wide variety of garments with diverse texture patterns. In addition, we also provide results on the VITON-HD dataset.

D. Discussion to compare the chosen works

1) Comparative Analysis:

- **Resolution and Image Quality**: VITON-HD excels in generating high-resolution images, crucial for applications where image detail and quality are paramount. In contrast, the other two approaches focus more on flexibility and handling unpaired data (Image Based Virtual Try-On Network from Unpaired Data) or diverse body shapes and poses (TryOnDiffusion).
- **Dataset Requirements**: The Image Based Virtual Try-On Network from Unpaired Data stands out for its minimal dataset requirements, only needing single images of garments and people, unlike the extensive paired datasets required by VITON-HD and TryOnDiffusion.
- **Technical Complexity**: TryOnDiffusion presents a more complex architecture with its Parallel-UNet and cross-attention mechanisms, aiming to address more challenging variations in body shapes and poses. In comparison, VITON-HD and Image Based Virtual Try-On Network from Unpaired Data use more straightforward yet effective deep learning techniques.
- **Application and Scalability**: The Image Based Virtual Try-On Network from Unpaired Data is highly scalable and practical due to its use of unpaired images, making it suitable for real-world applications with fewer data constraints. VITON-HD, while producing high-quality images, might be limited by its reliance on high-resolution paired datasets. TryOnDiffusion offers a robust solution for handling diverse and complex scenarios, making it highly adaptable but also demanding in terms of computational resources.
- **Conclusion**: Each paper presents unique strengths tailored to specific aspects of the virtual try-on challenge. VITON-HD excels in high-resolution synthesis, the Image Based Virtual Try-On Network from Unpaired Data offers practical scalability, and TryOnDiffusion provides a robust approach to handling diverse body shapes and poses. These differences highlight the various directions and innovations in the field of virtual try-on technology, addressing a wide range of challenges and applications.

III. Review 1 : Image Based Virtual Try-On Network from Unpaired Data

A. Summary & Contributions

The paper presents Outfit-VITON, a novel virtual try-on method that uses single images of garments and people to generate realistic try-on images. This approach eliminates the need for paired datasets or 3D models, making it scalable and practical. The method combines shape and appearance generation networks to produce a cohesive outfit from multiple garments, with an online optimization step for enhanced detail.

B. Strengths and Weaknesses

1) Originality:

- **Strengths**: The approach is original in its use of unpaired data and its ability to synthesize complete outfits from multiple garments. The incorporation of online optimization for detail refinement is also innovative. Collecting and training data is much easier than gathering 3D data or paired training images across multiple scales.
- **Weaknesses**: While the method is novel, it builds upon existing GAN and autoencoder techniques, and its innovation lies more in the application and combination of these techniques rather than in fundamentally new algorithms.

2) Quality:

- **Strengths**: The method demonstrates high-quality results both quantitatively and qualitatively, showing superior performance compared to existing methods in handling detailed garments. It introduces real-time optimization capabilities to accurately synthesize fine details such as textures and logos in the final outfit.
- **Weaknesses**: The evaluation is thorough but could benefit from more diverse datasets to test the model's robustness across different body types and garment styles.

3) Clarity:

- **Strengths**: The paper is well-written, with clear explanations of the methodology and detailed descriptions of the networks used.
- **Weaknesses**: Some technical details, particularly around the training process and parameter choices, could be expanded for greater clarity.

4) Significance:

- **Strengths**: The work has significant practical implications for online retail, offering a scalable and user-friendly solution to virtual try-on that could enhance customer experience and reduce return rates. By integrating multiple garments into a cohesive outfit, it provides an enhanced virtual fitting experience, and allows the user to control the final outfit they wear.
- **Weaknesses**: The significance could be further validated by user studies or real-world implementation trials.

C. Questions & Suggestions

- How does the system handle extreme variations in body shapes and sizes?
- Can the model be adapted to handle accessories such as hats and shoes in addition to clothing?
- What are the computational requirements for the online optimization step, and how does it impact real-time performance?

D. Limitations and Potential Negative Societal Impact

- **Limitations**:
  - The method may struggle with extreme variations in body shapes and sizes not well-represented in the training data.
  - Real-time performance might be limited by the computational demands of the online optimization step.
  - The quality of the output heavily relies on the quality of the input images and segmentation accuracy.
- **Potential Negative Societal Impact**:
  - There is a risk of reinforcing unrealistic body image standards if the model fails to accurately represent diverse body types.
  - Misrepresentation of garment fit and appearance could lead to customer dissatisfaction and increased returns.
- **Constructive Suggestions for Improvement**
  - Incorporate a diverse range of body shapes and sizes in the training data to improve the model's robustness.
  - Optimize the online refinement process to reduce computational load and enhance real-time applicability.

IV. Review 2 : VITON-HD

A. Summary & Contributions

One of the limitations of existing Virtual Try-On models is that they cannot produce high-resolution images. As the resolution increases, artifacts from misaligned parts become more prominent. Additionally, the architectures used in existing methods perform poorly in generating high-quality body parts and maintaining the sharpness of the clothing texture.

To address these issues, the authors of this paper developed a model called VITON-HD.

The paper introduces VITON-HD, a high-resolution virtual try-on system that incorporates several innovative components. These include a clothing-agnostic person representation, which eliminates the model's dependency on the original clothing worn by the person, and ALIAS normalization along with the ALIAS generator, which together address misalignment issues. VITON-HD is capable of producing high-quality synthetic images of people wearing various clothing items. Additionally, a new dataset was compiled to evaluate the performance of VITON-HD.

Photo-realistic 1024×768 virtual try-on 이미지를 합성하는 VITON-HD를 제안했다. 제안한 ALIAS normalization은 misaligned된 영역을 적절히 처리하고, 다중 스케일 정제를 통해 옷의 디테일을 보존하는 ALIAS 생성기 전체에 시맨틱 정보를 전파할 수 있다. 정성적, 정량적 실험을 통해 VITON-HD가 기존의 virtual try-on 방식을 큰 차이로 능가함을 입증했다.

B. Strengths and Weaknesses

1) Originality:

- **Strengths**: The paper is original in its introduction of ALIAS normalization and the clothing-agnostic person representation, which significantly improve image quality and resolution. The novel approach to handling misalignment and the detailed synthesis of high-resolution

virtual try-on images set this work apart from prior research.

- **Weaknesses**: While the method is innovative in its application, it builds upon existing deep learning frameworks such as GANs and autoencoders. The novelty primarily lies in the combination and application of these techniques rather than the development of entirely new algorithms.

2) Quality:

- **Strengths**: The experiments demonstrate the method's superior performance compared to existing approaches, showcasing high-quality results both quantitatively and qualitatively. The use of a high-resolution dataset and the comprehensive evaluation of the method highlight its effectiveness in generating realistic virtual try-on images.
- **Weaknesses**: The evaluation, though thorough, could benefit from more diverse datasets to test the model's robustness across different body types, garment styles, and varying conditions. This would ensure the method's generalizability and practical applicability in real-world scenarios.

3) Clarity:

- **Strengths**: The paper is well-written, with clear explanations of the methodology and detailed descriptions of the networks used. Visual aids, such as diagrams and example images, effectively support the text, making complex concepts accessible to readers.
- **Weaknesses**: Some technical details, particularly around the training process and parameter choices, could be expanded for greater clarity. Further elaboration on the implementation of ALIAS normalization and its impact on the results would enhance the reader's understanding.

4) Significance:

- **Strengths**: The advancements made by VITON-HD are significant for the field of virtual try-on, addressing key limitations of previous methods. The ability to produce high-resolution, photo-realistic images has substantial practical implications for online retail, potentially enhancing customer experience and reducing return rates.
- **Weaknesses**: The significance could be further validated by user studies or real-world implementation trials. Additionally, exploring the societal and ethical implications of virtual try-on technology, including privacy concerns and the potential for misuse, would strengthen the overall impact and relevance of the work.

C. Questions & Suggestions

- How does the model handle varying body poses and complex clothing patterns?
- Can the ALIAS normalization be applied to other image synthesis tasks?
- What are the computational requirements for training and inference?
- The ALIAS generator is said to fill misaligned areas with clothing texture. Does this mean that the shape of the clothing changes?

D. Limitations and Potential Negative Societal Impact

The primary limitation of VITON-HD is its dependency on high-quality segmentation maps and pose estimations, which may not always be accurate. This reliance can lead to errors in the final virtual try-on images if the initial segmentation or pose estimation is flawed.

Figure 15: Failure cases of VITON-HD.

Fig. 6. Failure cases of VITON-HD

위 이미지는 segmentation 맵이 부정확하게 예측되거나 옷깃 안쪽 영역이 다른 의류 영역과 구분되지 않아 발생한 모델의 실패 사례를 보여준다.

이 모델의 한계는 다음과 같다. VITON-HD는 아래쪽 의류 항목을 보존하도록 학습되어 대상 의류의 표현(예: 집어넣었는지 여부)이 제한된다. 다음으로, 데이터 세트는 대부분 날씬한 여성과 상의 이미지로 구성되어 있기 때문에 추론 과정에서 VITON-HD는 제한된 범위의 체형과 의상만 처리할 수 있다. 마지막으로, VITON-HD를 포함한 기존의 virtual try-on 방식은 실제 착용 이미지에 대한 강력한 성능을 제공하지 못한다.

Potential negative societal impacts include privacy concerns and the misuse of virtual try-on technology for creating deceptive images. The creation and distribution of manipulated images can lead to issues such as false advertising, identity theft, and other forms of digital deception.

Constructive suggestions for improvement include:

1) **Improving Robustness**: Enhance the robustness of the segmentation and pose estimation components to ensure more accurate and reliable outputs, even when working with lower-quality inputs.
2) **User Studies**: Conduct user studies and real-world implementation trials to gather feedback and validate the practical applicability and user acceptance of the technology.
3) **Diverse Datasets**: Utilize more diverse datasets to test the model's performance across different body types, garment styles, and various conditions, ensuring broader applicability and reducing bias.

V. Review 3 : TryOnDiffusion; A Tale of Two UNets

A. Summary & Contributions

"TryOnDiffusion: A Tale of Two UNets" by Luyang Zhu et al. proposes a novel approach to virtual apparel try-on, aiming to address the significant challenges in synthesizing photorealistic visualizations of garments on different body shapes and poses. The paper introduces a diffusion-based architecture called Parallel-UNet, which integrates two UNets that work together through a cross-attention mechanism. This architecture allows for implicit garment warping and blending in a single process, preserving garment details and accommodating substantial body pose and shape variations. The method is evaluated on a dataset of 4 million image pairs, achieving state-of-the-art performance both qualitatively and quantitatively.

B. Strengths and Weaknesses

Strengths:

- **Originality**: The integration of two UNets with a cross-attention mechanism is a novel approach in the virtual try-on domain. The implicit warping strategy without explicit pixel displacement estimation is innovative.
- **Quality**: The paper demonstrates robust experimental results, both qualitatively and quantitatively, showing significant improvements over state-of-the-art methods.
- **Clarity**: The paper is well-structured, with clear explanations of the methodology, experiments, and results. The visual examples provided effectively illustrate the improvements made by TryOnDiffusion.
- **Significance**: The method significantly advances the capabilities of virtual try-on technology, making it more applicable to real-world scenarios where pose and shape variations are common.

Weaknesses:

- **Complexity**: The architecture and training process are complex, potentially making it challenging for others to reproduce the results without significant computational resources and expertise.
- **Generalization**: While the model performs well on the dataset used for training, its performance on entirely unseen datasets or extreme edge cases is not extensively discussed.

C. Questions & Suggestions

- Data Augmentation: What types of data augmentation techniques were used during training, if any, to enhance the model's robustness?
- Real-world Applications: How does the model perform in a real-world online shopping scenario with varying image qualities and backgrounds?
- User Study Details: Can more details be provided about the user study, such as the demographic diversity of the participants and specific criteria used for ranking the images?
- 신경망에서 암시적 워핑을 어떻게 구현할 수 있는가?

D. Limitations and Potential Negative Societal Impact

Limitations:

- **Pose Estimation Errors**: The performance of the model heavily relies on accurate human parsing and pose estimation, which can be challenging in certain scenarios.
- 본 논문은 상반신 의상에 초점을 맞추었고 전신 시착은 실험하지 않았다.
- 의상에 구애받지 않는 RGB를 통해 정체성을 표현하는 것은 이상적이지 않다. 때로는 정체성의 일부만 보존할 수 있기 때문이다.
- 학습 및 테스트 데이터셋은 대부분 깨끗하고 균일한 배경을 가지고 있으므로 더 복잡한 배경에서 방법이 어떻게 수행되는지 알 수 없다.

Potential Negative Societal Impact:

- **Body Image Issues**: As with any virtual try-on technology, there is a potential risk of exacerbating body image issues if the technology is used to promote unrealistic or unhealthy body standards.
- **Privacy Concerns**: The use of personal images for virtual try-on could raise privacy concerns, particularly if data is not handled securely.

Suggestions for Improvement:

- **Efficiency Improvements**: Future work could focus on optimizing the computational efficiency of the model, making it more accessible for deployment in real-time applications.
- **Robustness Enhancements**: Enhancing the robustness of the model to handle a wider range of real-world scenarios, including different image qualities and backgrounds, would be beneficial.

VI. Future Work

A. The shortcomings or failure cases of the existing works

Computer Vision 분야에서 GAN 기반의 연구들이 Virtual Try-On 태스크의 성능 향상에 기여했음에도 불구하고, 상용 서비스에 접목 하기에는 다소 아쉬운 부분들이 있었습니다. 최근 Diffusion 기반의 Virtual Try-On 기술 연구 및 개발이 공개 됨에 따라 보다 Realistic한 결과가 나타나고 있습니다. 하지만 여전히 General 서비스 상용화를 위해 해결해야 할 숙제가 많이 남아있다고 판단하고 있습니다. 예를 들어 스포츠 영상에서의 인물들은 자세가 역동적인 경우가 많아 기존 SOTA 기술로 상용서비스를 커버하기 어려운 점이 있기 때문입니다.

B. Ideas for future extensions & New applications of VITON

Diffusion 모델 기반의 Virtual Try-On 네트워크에 대한 추가 학습을 통해 인물의 체형 혹은 의류 형태의 아티팩트를 최소할 수 있도록 고도화 개발에 집중합니다. Virtual Try-On 기술을 활용한 AI 숏폼 콘텐츠를 생성하여 상용화 할 수 있는 서비스도 기획할 수 있을 것입니다. 기존 쇼핑몰 사용자를 위한 서비스가 아닌 **디자이너를 위한 서비스**: 디자이너를 돕는 어시스턴트 서비스입니다. **디자이너가 옷을 스케치하고 스케치와 함께 옷에 대한 정보를 넣으면 그에 맞게 옷을 생성하고 그 옷을 모델에게 입혀보게 하는 것**입니다.

REFERENCES

[1] Assaf Neuberger, Eran Borenstein, Bar Hilleli, et al (2020) Image Based Virtual Try-on Network from Unpaired Data. CVPR 2020.

[2] Seunghwan Choi* Sunghyun Park* Minsoo Lee* Jaegul Choo, [KAIST, Daejeon, South Korea] (2021) VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. CVPR 2021.

[3] Luyang Zhu, Dawei Yang, Tyler Zhu, et al (2023) TryOnDiffusion: A Tale of Two UNets. CVPR 2023.

[4] Dan Song1, Xuanpu Zhang1, Juan Zhou1, et al (2023) Image-Based Virtual Try-On: A Survey.

[5] https://www.sktaifellowship.com/1bf551b2-17bf-467c-ae07-44c82019b528

[6] https://juniboy97.tistory.com/49