온디바이스 환경을 위한 Diffusion 기반 VTON 파이프라인 최적화

진경은, 신은호, 이유진, 전윤호*
국립한밭대학교
20227005@edu.hanbat.ac.kr, eunho@edu.hanbat.ac.kr, yujin@edu.hanbat.ac.kr *yhjeon@hanbat.ac.kr

Optimization of Diffusion-based Virtual Try-On Pipeline for On-device Deployment

Jin Gyeong Eun, Sin Eun Ho, Lee Yu Jin, Jeon Yun Ho*
Hanbat National University

요 약

본 논문은 Diffusion 기반 가상 착용(Virtual Try-On, VTON) 파이프라인을 온디바이스 환경에 최적화하는 방법을 제안한다. 기존 VTON 모델은 높은 계산 비용으로 인해 서버 기반 추론에 의존하며, 이 과정에서 사용자 이미지가 외부로 전송되어 개인정보 노출 위험이 존재한다. 이를 해결하기 위해 본 연구에서는 CatVTON을 기반으로 분할 모델(U-Net-R50)과 DPM++ 스케줄러를 적용한 파이프라인을 설계하였다. 또한 BG-20K 데이터셋을 활용한 배경 합성 기반 데이터 증강으로 실제 환경에서의 분할 성능을 향상시켰다. 실험 결과, 제안하는 파이프라인은 기존 대비 분할 성능(mIoU 0.9404)과 생성 품질(SSIM 0.6332)에서 우수한 성능을 보였으며, NVIDIA Jetson Orin Nano에서 약 3.75배 빠른 추론 속도를 달성하였다.

Ⅰ. 서 론

이미지 기반 가상 착용(Virtual Try-On, VTON) 기술은 사용자의 이미지와 의류를 합성하여 착용 모습을 제공하는 기술로, 온라인 쇼핑에서 반품률 감소에 기여할 것으로 기대된다[1][2]. 최근 Diffusion 기반 VTON 모델들이 높은 품질의 합성 이미지를 생성하며 주목받고 있으나, 반복적인 denoising 과정으로 인한 높은 계산 비용과 수십억 개의 파라미터로 인해 서버 기반 추론에 의존할 수밖에 없다. 이 과정에서 사용자 인물 이미지가 외부로 전송되면서 개인정보 노출 위험이 발생한다. 본 논문에서는 CatVTON 기반의 온디바이스 VTON 파이프라인을 제안한다. 제안 파이프라인은 NVIDIA Jetson Orin Nano에서 기존 대비 약 3.75배 빠른 추론 속도를 달성하면서 분할 성능(mIoU 0.9404)과 생성 품질(SSIM 0.6332)을 유지하였다.

Ⅱ. 관련된 연구

2.1 가상 착용(Virtual Try-On, VTON)

이미지 기반 가상 착용 기술은 사용자의 이미지와 선택한 의류를 자연스럽게 합성하여 실제 착용한 모습과 유사한 가상 이미지를 제공하는 기술이다. 최근에는 인페인팅 기반 Diffusion 모델을 가상 착용 문제에 맞게 파인튜닝하여 VTON 과제에 적용한 사례들이 주목받고 있다. 대표적으로 TryOnDiffusion[6], IDM-VTON[7] 등이 제안되었으며 이들은 의류의 디테일과 텍스처를 보존하면서 자연스러운 착용 이미지를 생성한다.

2.2 CatVTON

CatVTON[8]은 Diffusion 기반 VTON에서 흔히 사용되는 텍스트 인코더나 병렬 UNet 등을 최소화하여 구조를 단순화한 VTON 모델이다. 입력 조건을 하나의 입력으로 결합(Concatenation)해 이를 단일 UNet 구조로 처리한다는 점이 주요 특징이다. 이러한 설계는 파라미터 규모와 메모리 사용량을 감소시켜 온디바이스 환경에 비교적 유리하다.

Ⅲ. 제안하는 방법

3.1 최적화된 파이프라인

그림1은 제안하는 온디바이스 VTON 파이프라인의 전체 구조를 나타낸다. 파이프라인은 마스킹 단계와 생성 단계로 구성된다. 마스킹 단계에서는 입력된 인물 이미지로부터 의류 영역을 분리한다. 분할 네트워크 S_θ 는 인물 이미지를 입력받아 의류 영역에 해당하는 이진 마스크를 출력한다. 생성된 마스크는 원본 이미지와 element-wise 곱 연산을 통해 의류 영역이 제거된 인물 이미지를 생성한다.

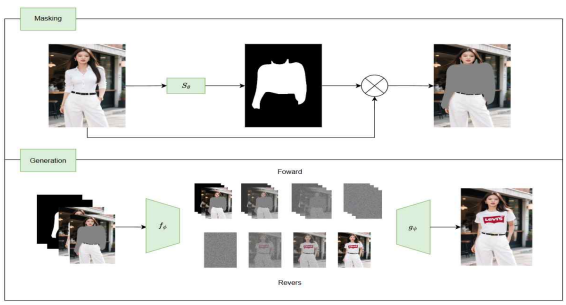

그림 1. 제안하는 VTON 파이프라인 구조

생성 단계에서는 마스킹된 인물 이미지와 목표 의류 이미지를 입력으로 받아 최종 착용 이미지를 생성한다. 두 입력은 인코더 f_ϕ 를 통해 잠재 벡

터로 매핑한 후 공간 차원에서 결합한다. 결합된 잠재 벡터는 확산 모델의 순방향 역방향 과정을 거치며 디코더 g_ψ를 통해 목표 의류가 착용된 최종 이미지로 복원한다.

3.2 데이터 증강

VITON-HD 데이터셋[3]으로 학습된 기존 분할 모델은 단순한 배경 이미지가 주를 이루는 학습 데이터 특성으로 인해 복잡한 배경이 존재하는 실제 추론 환경과의 도메인 갭이 발생한다(그림 2). 분할 마스크의 품질 저하는 후속 생성 결과의 품질 저하로 이어지므로 강건한 분할 성능 확보는 전체 파이프라인의 성능을 위해 필수적이다.

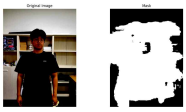

그림 2. 기존

그림 3. 증강

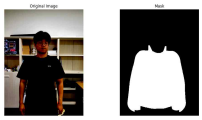

그림 4. 개선

이를 위해 BG-20K 데이터셋[4]을 활용하여 배경 다양성을 반영한 증강 학습 데이터셋(그림 3)을 구축하고, 이를 분할 모델 학습에 적용함으로써 배경 변화에 강건한 분할 성능을 확보한다(그림 4). 데이터 증강 파이프라인은 거대 비전 모델인 SAM[5]을 활용해 전경과 배경을 분할하고 배경을 합성하는 방식을 사용했다.

Ⅳ. 실험

4.1 분할 성능 비교

비교 대상인 AutoMasker는 CatVTON[8]에서 제안된 분할 파이프라인이다. 제안하는 U-Net-R50은 ResNet-50 인코더 기반 U-Net 구조이며, VITON-HD와 BG-20K를 활용한 14,684장의 증강 데이터로 학습하였다. 평가는 동일한 방식으로 배경 합성을 적용한 데이터셋에서 mIoU와 Dice score로 측정하였다.

Model	mIoU	Dice
AutoMasker[8]	0.7641	0.8635
U-Net-R50(Ours)	0.9404	0.9683

표1. 분할 평가결과

표 1에서 보듯이, 제안하는 U-Net-R50은 AutoMasker 대비 mIoU 0.18, Dice 0.10 높은 성능을 보였다.

4.2 생성결과

기존 파이프라인(AutoMasker + DDIM 50 step)과 제안하는 파이프라인(U-Net-R50 + DPM++ 15 step)의 성능을 비교하기 위해 평가 데이터셋에서 100장을 샘플링하여 5회 반복 실험을 수행하였다. 실험은 NVIDIA RTX A6000 Ada GPU에서 진행하였으며, 생성 품질 지표로 SSIM을 사용하고, 처리 효율성 평가를 위해 전체의 추론 시간을 함께 측정하였다.

Method	SSIM	Time
CatVTON Pipeline	0.6042 ± 0.078	3m 11s ± 1.5s
Ours	0.6332 ± 0.093	1m ± 0.5s

표2. 파이프라인 평가결과

표 2에서 보듯이, 제안하는 파이프라인은 SSIM 0.6332로 기존(0.6042) 대비 높은 생성 품질을 보였으며, 추론 시간도 약 3배 단축되었다. 이는 마스크 품질이 최종 생성 결과에 주요한 영향을 미친다는 점을 보여준다.

4.3 Jetson Orin Nano 속도 실험

제안하는 파이프라인의 엣지 디바이스 적용 가능성을 검증하기 위해 NVIDIA Jetson Orin Nano 환경에서 추론 속도를 측정하였다. 분할 모델만을 대상으로 단일 이미지 추론 시간을 비교하였다.

Method	Time/Img
CatVTON Pipeline	30s
Ours	8s

표3. Jetson Orin Nano 평가결과

표 3에서 보듯이, 기존 파이프라인은 30초가 소요된 반면, 제안하는 파이프라인은 8초로 약 3.75배 빠른 속도를 달성하였다. 이는 제안하는 파이프라인이 엣지 디바이스 환경에서도 효율적으로 동작함을 보여준다.

Ⅴ. 결론

본 논문에서는 Diffusion 기반 가상 착용 파이프라인을 온디바이스 환경에 최적화하는 방법을 제안하였다. U-Net-R50 분할 모델과 DPM++ 스케줄러를 적용하여 기존 CatVTON 파이프라인 대비 생성 품질과 추론 속도를 동시에 개선하였다. 또한 BG-20K 기반 배경 합성 데이터 증강을 통해 복잡한 배경 환경에서도 강건한 분할 성능을 확보하였다. 실험 결과, 제안하는 파이프라인은 mIoU 0.9404, SSIM 0.6332를 달성하였으며, NVIDIA Jetson Orin Nano에서 기존 대비 약 3.75배 빠른 추론 속도를 보였다.

ACKNOWLEDGMENT

"본 연구는 2025년 과학기술정통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음" (2022-0-01068)

참 고 문 헌

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An Image-based Virtual Try-on Network," in Proc. CVPR, 2018.

[2] H. Yang and W. Yu, "The role of virtual try-on technology in online purchase decision from consumers' aspect," Ind. Manag. Data Syst., vol. 119, no. 2, 2019.

[3] S. Choi, S. Park, S. Lee, and J. Choo, "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization," in Proc. CVPR, 2021.

[4] J. Li, J. Zhang, and D. Tao, "BG-20K: A Large-scale Dataset for Background Matting," in Proc. IJCAI, 2022.

[5] A. Kirillov et al., "Segment Anything," in Proc. ICCV, 2023.

[6] L. Zhu et al., "TryOnDiffusion: A Tale of Two UNets," in Proc. CVPR, 2023.

[7] S. Choi, S. Park, M. Lee, J. Park, and J. Choo, "IDM-VTON: Improving Diffusion Models for Authentic Virtual Try-On in the Wild," in Proc. ECCV, 2024.

[8] Z. Chong et al., "CatVTON: Concatenation is All You Need for Virtual Try-On with Diffusion Models," arXiv 2024