# Openpose 데이터 기반 StableVITON 모델 학습 시간 개선 연구

임수정*, 피하영, 김수영

*인천대학교 컴퓨터공학부, 성신여자대학교 수학과, 동국대학교 응용통계학과

lsj69311@gmail.com, kjy1ts@gmail.com, suu00k1459@gmail.com

# Study on Improving Training Time of StableVITON Model Using Openpose Data

Soo Jeong Lim*, Ha Yeong Pi, Su Young Kim

*Dept of. Computer Science and Engineering, Incheon National Univ.
Dept of. Mathematics, Sungshin Women's Univ., Dept of. Applied Statistics, Dongguk Univ.

요 약

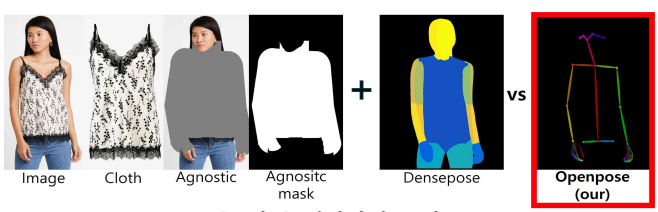
본 논문은 의류 이미지를 인물 이미지에 자연스럽게 입히는 것을 목표로 하는 이미지 기반 가상 착용(VITON) 분야에서, Stable Diffusion[5] 모델을 기반으로 한 Stable VITON[1] 모델을 사용하여 포즈 데이터 차이에 따른 학습 과정을 분석한다. 기존 연구들에서는 포즈 데이터로 주로 인체 표면을 매핑하는 복잡한 Densepose를 활용했지만, ControlNet[3]에서 성능이 입증된 인체 관절 위치 중심의 간단한 Openpose[2]를 활용하여, 의류 데이터 중심의 효율적인 학습이 가능한지 확인하였다. 그 결과, Openpose 데이터셋 기반 모델이 더 적은 학습 시간 내에 로고 중심의 의류를 정확하게 생성하였으며, 패턴과 프린팅 중심의 의류도 빠르게 학습하여 전반적인 디테일을 생성하는 것을 확인할 수 있었다. 이를 통해 VITON 분야에서 포즈 데이터에 따른 학습 시간과 성능에 대한 방향성을 제시하고자 한다.

## Ⅰ. 서 론

이미지 기반 가상 착용 모델(VITON:An Image-Based Virtual Try-On Network)은 의류 이미지를 인물 이미지에 자연스럽게 입히는 것을 목표로 한다. 최신 VITON 모델 중 하나인 StableVITON[1]은 Stable Diffusion[5] 모델을 기반으로 하며, 기존 VITON 모델들처럼 포즈 데이터로 인체 표면을 매핑하는 복잡한 Densepose를 사용해 왔다. 그러나 이러한 데이터셋 구성은 인체의 모든 표면을 매핑하는 복잡성 때문에 모델의 계산 비용을 증가시키고, 학습 과정에서 의류의 세부 정보를 정확하게 반영하는 데 많은 시간이 소요될 가능성이 있어 보인다(StableVITON: 4□A100 GPU로 약 100시간 학습, ViViD: 4□A100 GPU로 약 120시간 학습 etc). 2023년에 발표된 Diffusion 모델 기반의 ControlNet[3]에서는 인체의 주요 관절 위치만을 간결하게 표현하는 Openpose[2]를 사용하여 포즈 관련 정보를 효과적으로 표현하였기에, 이를 활용하여 Stable Diffusion[5] 기반 VITON 모델의 학습 시간을 개선하고자 한다.

## Ⅱ. 본론

### 1. 데이터셋




[그림1] 데이터셋 구성

본 연구에서는 VITON-HD[4] 데이터셋을 기반으로 실험을 진행하였다. 기존 모델들은 인물 이미지(Image), 의류 이미지(Cloth), 의류 정보를 제거한 인물 이미지(Agnostic), Agnostic의 마스크(Agnostic mask), 인체 표면을 맵핑하여 포즈를 표현하는 Densepose를 포함하였다. 그러나 본 연구에서는 동일한 VITON-HD 데이터셋을 사용하되, 포즈 표현 방식으로 Densepose 대신 인체의 주요 관절 위치만을 간결하게 표현하는 Openpose로 대체하여 학습을 수행하였다.

### 2. 기본 모델 구조 및 학습 설정

본 연구에서는 StableVITON[1]을 기본 모델로 사용하였다, 이 모델은 Stable Diffusion[5]을 기반으로 Zero Cross-Attention Block을 활용해 의류의 디테일한 특성을 인체 이미지에 정확하게 매핑하는 우수한 성능을 보여주었기에 선정되었다. 특히, 포즈 데이터(Densepose vs Openpose)에 따른 학습 시간의 차이를 비교하기 위해, 동일한 StableVITON 모델 구조에서 포즈 데이터만 다르게 설정한 두 데이터셋을 각각 1□A100 GPU로 500epochs(약 50시간)씩 학습하고, 그 과정을 독립적으로 관찰하였다.

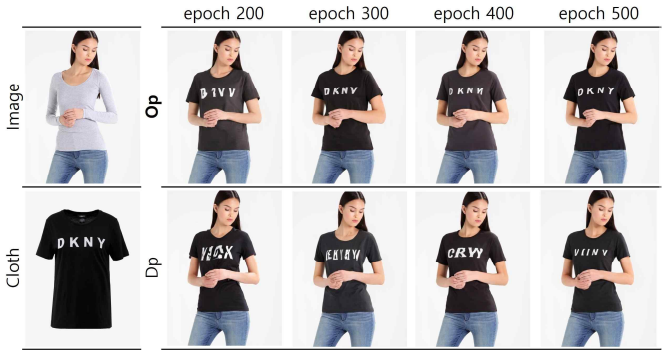
| Model | Dp | **Op (our)** |
|---|---|---|
| 정의 | Densepose 데이터셋 기반 StableVITON 모델 | Openpose 데이터셋 기반 StableVITON 모델 |

[표1] Dp, Op(our) 모델의 정의 및 관계

### 3. 실험 결과

학습의 과정을 관찰한 결과, Op 모델이 Dp 모델에 비해 의류의 로고, 패턴, 프린팅과 같은 복잡한 세부 정보를 더 적은 시간 내에 학습하는 성능을 보였다. 또한 이는 Densepose 데이터셋을 기반으로 한 StableVITON [1] 최종 모델과 비교 해도 유사한 성능을 보이고 있다.

### 1) 로고 중심의 의류



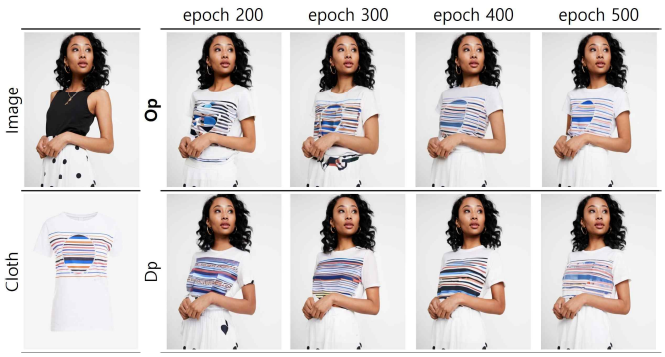

[그림2] 로고 중심의 의류에 대한 학습 진행 비교

[그림2]를 보면, Op 모델은 epoch 200 초반에 "DKNY" 로고 윤곽을 생성하기 시작해, epoch 500 이후에는 거의 완벽하게 생성하여 세부 정보를 선명하게 반영하였다. 반면, Dp 모델은 epoch 500에서도 유사한 로고조차 생성하지 못하였다.


[그림3] 로고 중심의 의류에 대한 생성 결과 비교

[그림3]에서는 Op 모델이 더 적은 학습 시간 내에 StableVITON 최종 모델과 유사한 이미지를 생성한 결과를 보여준다.

### 2) 패턴 중심의 의류




[그림4] 패턴 중심의 의류에 대한 학습 진행 비교

[그림4]를 보면, Op 모델은 epoch 200 초반부터 줄무늬 패턴 내 타원 무늬의 윤곽을 인식하기 시작했으나, Dp 모델은 epoch 500에 이르러서도 대략적인 색상만 생성했을 뿐, 타원 무늬를 제대로 생성하지 못하였다.


[그림5] 패턴 중심의 의류에 대한 생성 결과 비교

[그림5]에서는 StableVITON 최종 모델과 Op 모델 모두 타원 무늬와 전반적인 색상을 잘 생성하였지만, 더 많은 학습 시간을 사용한 StableVITON 최종 모델이 원본 의류의 색상을 더 선명하게 생성하였다.

### 3) 프린팅 중심의 의류


[그림6] 프린팅 중심의 의류에 대한 학습 진행 비교

[그림6]에서는 Op 모델은 epoch 200 초반에 의류 프린팅의 전체적인 외형을 생성하였으나, Dp 모델은 기본적인 형태조차 생성하지 못하였다.


[그림7] 프린팅 중심의 의류에 대한 생성 결과 비교

[그림7]에서는 StableVITON 최종 모델과 Op 모델 모두 프린팅의 전반적인 특성을 잘 생성하였지만, 더 많은 학습 시간을 사용한 StableVITON 최종 모델의 결과가 프린팅의 세밀한 테두리까지 생성하였다.

## Ⅲ. 결론

본 연구에서는 Stable Diffusion 기반의 StableVITON 모델에서 Openpose라는 간결한 포즈 데이터를 활용하여, 복잡한 의류 세부 정보에 대한 가상 착용 시스템의 학습 시간과 비용을 절감할 수 있음을 확인하였다. 그러나 매우 세밀한 색상과 디자인 표현에서는 기존 StableVITON 모델의 최종 결과에 비해 디테일이 다소 부족했다. 이러한 점을 고려하여 다양한 데이터셋에 따른 모델의 학습 시간과 성능을 심도 있게 분석하는 후속 연구가 필요하다. 이를 통해 데이터셋별 학습 시간 개선을 위한 효율적인 모델 학습 기준을 마련할 수 있을 것으로 기대한다.

## 참 고 문 헌

[1] 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.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.,* 2019.

[3] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffustion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.,* pp. 3836-3847, 2023.

[4] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* pp. 14131-14140, 2021.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* pp. 10684-10695, 2022.