「정보시스템연구」 제31권 제4호
한국정보시스템학회
2022년 12월, pp. 145～160

http://dx.doi.org/10.5859/KAIS.2022.31.4.145

딥러닝 의류 가상 합성 모델 연구: 가중치 공유 & 학습 최적화 기반 HR-VITON 기법 활용*

이현상** · 오세환*** · 하성호****

〈목 차〉

Ⅰ. 서 론

최근 경영정보시스템 분야에서 딥러닝 시스템을 기반으로 특정 산업의 문제점을 해결하고자 하는 연구들이 시도되고 있다(조보근 외., 2020; 원종관 외., 2021; 신은경 외., 2021). 특히, 패션 산업 E-Commerce에서는 온라인 거래 사용자의 신체 치수와 의복 크기가 부합하지 않아서 사용자의 서비스 만족도가 낮아지거나, 의류 판매자의 입장에서 큰 손실이 발생할 수 있는 문제점이 존재한다. 최근에는 이러한 문제를 해결하기 위해서 패션 분야 기업들은 AI 기반 의류 가상 피팅 시스템을 도입하고 있다. 2020 Adobe Sneaks 세션에서는 Zara의 의류 가상 합성 시스템을 소개했다(Ayush Chopra, 2020). ZARA의 경우 매년 약 1만 2천 개의 다양한 디자인을 출시하여 지속적으로 패션 트렌드에 부합하는 제품을 소비자에게 제공하는 판매 전략을 가지고 있다. ZARA는 다양한 디자인을 출시하기 때문에 각기 체형이 다른 모델들에게 촬영할 시 막대한 시간과 비용이 소모된다는 문제점을 가지고 있다. 이와 같은 문제

* 이 논문은 2022년도 산업통상자원부 산업혁신기반구축사업 재원으로 수행된 연구임(P114000015).
** 경북대학교 경영학부 박사과정, coolwin200@gmail.com(주저자)
*** 경북대학교 경영학부 부교수, sehwano@knu.ac.kr
**** 경북대학교 경영학부 교수, hsh@knu.ac.kr(교신저자)

로 인해 ZARA는 Adobe의 엔지니어링 팀과 협업하여 딥러닝, 3D 모델링, 사진 자동 편집 등의 기술을 활용하여 의류 가상 합성 시스템을 개발했다.

하지만 3D 모델링 기술이 같이 적용되어야 하는 의류 가상 합성 기법은 착용자 및 의복의 정확한 3D 치환이 필요하다는 점에서 한계점이 있다. 3D 모델링은 고도화된 기술이 필요하며 이에 시간과 비용이 추가적으로 발생할 수 있다. 따라서 기존 연구에서는 의류 제품 사진과 착용자를 바로 합성할 수 있는 딥러닝 기술을 연구하고 있다(Han et al., 2018; Lyu et al., 2021). 의류를 가상으로 합성할 수 있는 VITON 기법은 이미지 세그멘테이션(segmentation) 및 포즈 추정(pose estimation), GAN(Generative Adversarial Network) 기반 이미지 생성 등 고도의 딥러닝 기술이 요구된다(Han et al., 2018; Yang et al., 2020; Lyu et al., 2021; Lin et al., 2022; Lee et al., 2022). 하나의 시스템에서 다양한 딥러닝 모델의 추론 과정이 필요하기 때문에 모델 학습에 어려움이 존재한다. 최신 논문에서는 고해상도 이미지를 합성하기 위해서 다양한 모델 학습 방법론을 연구하고 있다(Yang et al., 2020; Lyu et al., 2021; Lin et al., 2022; Lee et al., 2022).

그러나 기존 연구의 경우 세 가지 한계점이 존재한다. 첫 번째, 기존 연구에서는 정면 제품 및 착용 이미지만을 활용하여 딥러닝 모델을 학습했다. 실제 패션 커머스에서는 정면뿐만 아니라 측면, 후면, 혹은 유동적인 자세에 따른 이미지가 필요하며, 서비스 사용자의 경우에도 다양한 측면에서의 의류 착용 정보가 필요할 수 있다. 두 번째, 이미지 생성 모델의 특성상 학습에 필요한 시간과 비용 및 자원의 수준이 매우 높다는 것이다. 이에 민간 차원의 연구가 어렵고, 고품질의 모델 추론 결과를 도출하기 위한 시간과 학습 자원 및 비용이 크게 발생한다는 문제점이 있다. 세 번째, 기존 연구에서는 착용자의 체형을 고려하기 위해서 고스트컷(ghostcut) 제품 사진을 활용한다. 고스트컷이란 패션 제품의 착용 시 형태를 보기 위해서 마네킹을 활용하여 제품 사진을 보정한 것을 의미한다. 서비스 환경을 고려했을 때, 이러한 부분에서도 제품 단위의 작업이 수행되어야 하기

<그림 1> 좌 – 일반 제품 사진, 우 – 고스트컷 제품 사진

때문에 시간과 비용이 추가로 발생할 수 있다.

이에 따라서 본 연구의 목적은 정면뿐만 아니라 후면 제품 및 착용 사진을 활용하여 추가적인 자세에서의 딥러닝 모델 활용 가능성을 검토하는 것이다. 본 연구에서는 따로 정면과 후면 이미지 데이터를 따로 학습하는 방식이 아니라, 정면 및 후면 모델 네트워크의 가중치를 공유하는 방식을 활용했다. 또한 MP-GA (Mixed Precision-Gradient Accumulated) 기법을 활용하여 모델 학습에 소요되는 시간과 비용을 절감하고자 한다. 최종적으로 기존 연구와 같은 고스트컷 제품 이미지가 아닌 일반적인 패션 제품 이미지를 활용하여 딥러닝 모델을 실험했다. 본 연구를 통해서 국내에서의 제품 및 착용 이미지를 활용한 딥러닝 의류 가상 착용 모델 연구에 대해서 실증하고, 국내 패션 커머스에도 적용될 수 있도록 해당 분야의 연구가 활성화될 수 있기를 기대한다. 특히, 국내에서도 섬유소재산업 빅데이터 플랫폼 구축 사업을 통해서 온라인 섬유 거래를 활성화하고자 하는데, 해당 빅데이터 플랫폼에서 본 연구의 딥러닝 시스템을 활용할 수 있기를 기대한다 (매일신문, 2020).

Ⅱ. 관련 연구

딥러닝 기반 의류 가상 합성 모델 연구는 VITON(Virtual Try-On Network) 기법이 연구되면서 시작되었다(Han et al., 2018). VTION은 의복 및 착용자의 이미지가 주어졌을 때, GAN(Generative Adversarial Networks)을 기반으로 새로운 의복의 이미지를 착용자에게 합성할 수 있는 방법론이다. 실제로는 착용하지 않은 의류 제품을 가상으로 합성하기 위해서, VITON에서는 우선 착용자의 신체 정보를 데이터로 표현한다. 즉, 사람의 포즈 히트맵(pose heatmap), 몸 신체 영역과 머리 및 모발 영역을 독립적인 딥러닝 모델을 통해서 추론한다. 포즈 히트맵이란 사람의 관절 포인트를 나타내는 것으로 자세, 동작, 구조 등의 정보를 표현할 수 있다. VITON 연구에서는 신체 부위 집중 영역(PAF: part affinity fields) 기법을 활용한 다중 2d 포즈 추정 모델을 활용하여 포즈 히트맵을 추론했다(Cao et al., 2017). PAF란 비파라미터 표현(nonparameter represenatation) 기법을 통해서 사람과 신체를 연결하는 이미지 데이터의 특성을 추출하는 방식이다. PAF를 통해서 이미지의 글로벌 컨텍스트를 학습할 수 있기 때문에 사람 수가 많을 때에도 안정적인 추론 결과를 도출한다. 신체 영역의 경우 SSL (Self-supervised Structure-sensitive Learning) 모델을 활용하여 세그멘테이션을 수행했다 (Gong et al., 2017). SSL이란 대량의 고수준 글로벌 컨텍스트를 자기지도학습하는 방식으로, 사람 신체의 부위별 영역을 효과적으로 추출할 수 있는 딥러닝 알고리즘이다. Gong et al. (2017)는 기존의 신체 영역을 구분하는 벤치마크 데이터셋에서 확장성 및 다양성을 고려한 새로운 데이터셋 LIP(Look into Person)를 제시하기도 했다. 머리 및 모발 영역의 추론에서도 Gong et al. (2017)의 모델 결과물을 활용했다. 최종적으로 의복 대체 영역(clothing-agnostic)을 다양한 자세의 사람 이미지를 생성할 수 있는 딥러닝 기법과 구조적 일관성(structural coherence) 기반 의복 합성 시스템을 통해서 생

성한다(Liu et al., 2016; Zhe et al., 2017).

VITON 네트워크의 인코더 및 디코더 제너레이터(generator) 구조에서는 사람의 신체 데이터와 의복 제품 이미지 데이터를 학습하여 의복 대체 영역 및 실제 착용 이미지를 생성한다. 의복 대체 영역을 추론할 때는 L1 Loss, 실제 착용 이미지를 합성할 때는 Perceptual Loss를 통해서 딥러닝 모델의 가중치를 갱신한다. 정제 네트워크(refinement network)에서는 AI가 생성한 이미지를 후처리할 수 있는 TPS (Thin Plate Spline) 변환과 복합 학습 기법(Learn to Composite)을 활용한다. 여기서 TPS 기법을 통해 의복 대체 영역에 맞춰서 실제 제품 이미지를 워핑시키는데, 이 제품 워핑 이미지와 AI가 생성한 합성 이미지를 가지고 최종 의복 합성 이미지를 추론한다. 정제 네트워크의 최종 이미지 추론에도 Perceptual Loss가 적용되었다.

VITON 기법의 등장 이후 다양한 AI 기반 의복 가상 합성 모델 관련 연구가 시작되었다. CP-VTON(Characteristic-Preserving Virtual Try-On Network)을 개발한 Wang et al. (2018)은 기존 VITON 모델이 합성된 가상 의류 이미지가 질감, 로고, 자수 등의 특징적인 표현을 제대로 수행하지 못한다는 점을 지적했다. 의복 워핑 이미지 생성 시 세부적인 부분들이 제대로 표현할 수 없다는 것이다. CP-VTON 기법은 TPS가 아닌 GMM(Geometric Matching Module)이라는 새로운 워핑 이미지 생성 방식을 제시함으로써 합성 이미지의 특징적인 부분을 보완하고자 했다. Yang et al. (2020)은 기존 연구의 특징적인 부분까지 자연스럽게 구현하고자 했던 시도에 대해서 여전히 합성 이미지의 품질이 낮은 수준이라는 점을 지적했다. 이에 따라 ACGPN(Adaptive Content Generating and Preserving Network)이라는 새로운 기법을 제시했다. ACGPN은 합성 이미지의 의미론적 구조를 예측한 다음 의복 제품 이미지를 학습하는 방식을 활용했다. 여기서 의미론적 구조를 활용하는 것은 의복 제품의 셔츠, 재킷의 레이블링 및 착용자의 신체 정보를 모델에 반영하기 때문에 안정적으로 착용 이미지를 추론할 수 있다는 이점을 가지고 있다.

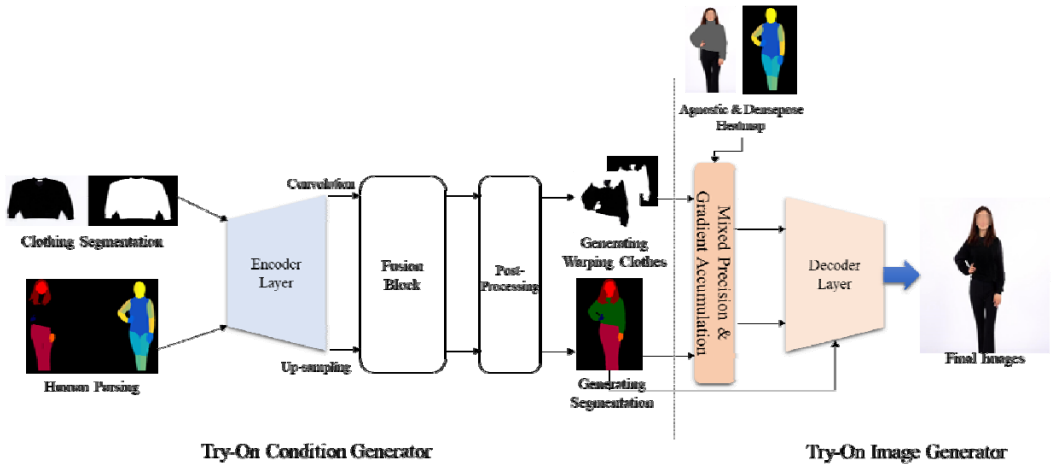
최근 연구 동향으로 Lin et al. (2022)은 PB(Parser-Based) 방식의 한계점을 지적하고 RMGN(Regional Mask Guided Network)이라는 새로운 PF(Parser-Free) 방식을 제안했다. PB란 사람의 신체 영역, 관절 포인트 구조 등의 데이터를 합성하는 방식으로, Lin et al. (2022)에 의하면 PB 기법은 원천 의복 이미지의 노이즈가 생성 이미지의 품질을 해칠 수 있다고 주장했다. 이에 따라 RMGN은 의복 이미지와 착용자의 데이터 특성을 명시적으로 융합할 수 있는 Regional Mask 방법론이 적용되었다. Lee et al. (2022)은 기존 VITON 기법이 의복 이미지를 착용자의 신체 영역에 맞춰 워핑하는 단계에서 모델 네트워크의 정보 교환이 없다는 문제점을 지적했다. 이에 따라 워핑 이미지와 신체 영역 사이에서 픽셀 압착 현상이 발생하여 생성 이미지의 품질이 낮아질 수 있다는 것이다. Lee et al. (2022)은 이러한 문제를 해결하기 위해 워핑 및 세그멘테이션 이미지 생성 단계의 통합 모듈을 활용했다.

Ⅲ. 제안 기법

3.1 HR-VITON

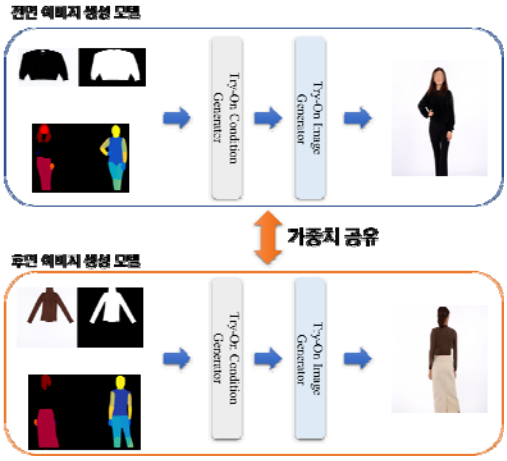
본 연구에서는 HR-VITON(High-Resolution Virtual Try-On) 기법을 활용하여 국내의 의복 제품 및 착용자의 이미지 데이터를 학습했다(Lee et al., 2022). HR-VITON 기법은 고해상도 이미지를 효과적으로 학습하기 위해서 기존 연구의 워핑 변환 기법을 사용하지 않는 통합 모듈 네트워크 구조를 제시했다. 이에 따라 HR-VITON은 워핑 및 세그멘테이션 생성의 기능을 가지고 있는 조건부 영역 생성기(condition generator)의 구조를 가지고 있다. 조건부 영역 생성기에서는 합성 이미지의 품질을 고려하여 잘못 합성된 이미지를 탐지하는 판별자(discriminator rejection)의 개념을 추가적으로 도입했다. 본 제안 기법은 <그림 2>와 같이 의복 조건부 영역 생성기(Try-On Condition Generator)와 의복 이미지 생성기(Try-On Image Generator) 모듈로 구성되어 있으며, 모델 규모가 큰 의복 이미지 생성기가 학습될 때 최적화 및 모델 일반화를 위한 MP-GA(Mixed Precision & Gradient Accumulation) 기법이 적용되었다.

<그림 2>의 HR-VITON 이미지 학습 프로세스는 다음과 같다. 첫 번째, 의복 대체 영역 및 사람의 신체 정보 데이터(Human Parsing)를 모델 인코더에 입력한다. 두 번째, 모델 인코더에서 학습한 의복 정보와 신체 정보를 통합(Fusion Block)하여 착용자 의복 영역을 생성(Generating Segmentation)한다. 세 번째, 두 번째 단계의 가중치를 컨볼루션(Convolution)하여 워핑된 옷 이미지와 마스킹(masking) 영역을 생성(Generating Warping Clothes)한다. 네 번째, 생성된 이미지들의 후처리를 위해서 의복 워핑 이미지에 사람의 신체 정보를 반영한다. 다섯 번째, 디코더 레이어에 워핑된 옷 이미지 및 영역, 착용자 의복 영역, 의복 대체 영역 및 Densepose 히트맵(Heatmap)을 모두 입력하여 최종 이미지(Final Images)를 생성한다. 본 연구에서는 딥러닝 의류 가상 합성 모델의 다양



<그림 2> MP-GA HR-VITON 네트워크 구조

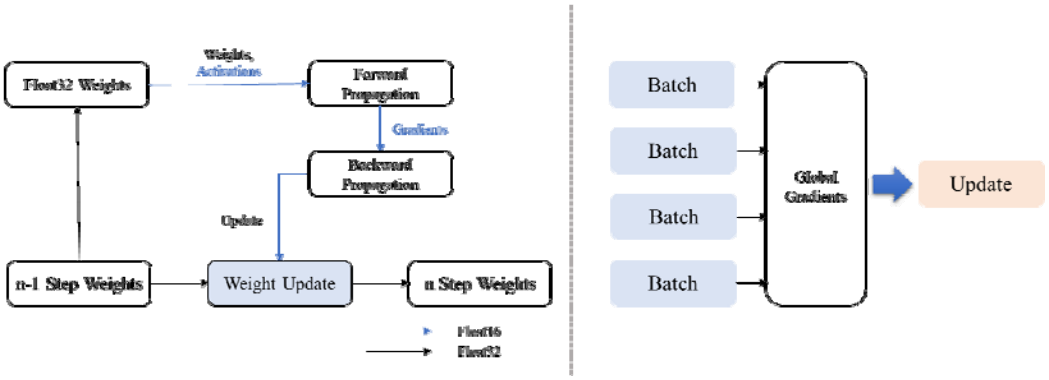
한 포즈를 고려할 수 있는지 실증하기 위해서 후면 의복 제품 이미지 및 착용자를 추가적으로 학습했으며, 가중치를 공유할 수 있도록 모델을 재설계했다. <그림 3>과 같이 가중치 공유 시 하나의 모델에서 전면 및 후면 이미지 데이터를 학습할 때 같은 손실 함수를 기준으로 가중치가 업데이트될 수 있도록 구성되었다.



<그림 3> 가중치 공유 모델 구조

3.2 MP-GA

기존 딥러닝 분야에서는 모델 네트워크의 복잡성, 데이터 및 가중치 수의 증가로 인한 학습 최적화 기법을 연구하고 있다(Ruder, 2016; Micikevicius et al,. 2017). 본 연구에서 사용된 MP(Mixed Precision) 기법은 일반적으로 학습에서 사용하는 Float32, 즉 부동소수점 32자리 수를 16자리로 축소시키는 방법론이다(Micikevicius et al., 2017). MP 방법론은 가중치 업데이트 단계에서 Float32를 유지하기 위해 손실함수를 보정하고, Float32 가중치를 계산하기 위해서 Float16 연산을 활용한다. 즉, 최종적으로 저장되는 가중치는 Float32 형태지만, 신경망 연산에서 Float16을 활용하는 방식이다. GA(Gradient Accumulation) 기법은 딥러닝 학습에서 그래디언트 값을 n 스텝만큼 누적시켜 한 번에 가중치를 업데이트하는 방식으로 배치(batch) 설정 시 메모리 문제를 상당 부분 보완할 수 있다(Ruder, 2016). 예를 들어, 딥러닝 학습에서 배치 크기를 32로 설정하고, GA 파라미터를 4로 설정했을 때, 실제 결과물에서 다소 차이가 있더라도 128의 배치 크기를 설정한 것과 근사한 결과 값을 도출할 수 있다는 것이다. 본 연구의 가상 의류 합성 피팅 모델은 가중치의 개수가 약 1.5억 개로 신경망의 복잡성이 크



<그림 4> Mixed Precision(좌) & Gradient Accumulation(우)

기 때문에 MP 및 GA 기법과 같은 방법론이 효과적일 수 있다.

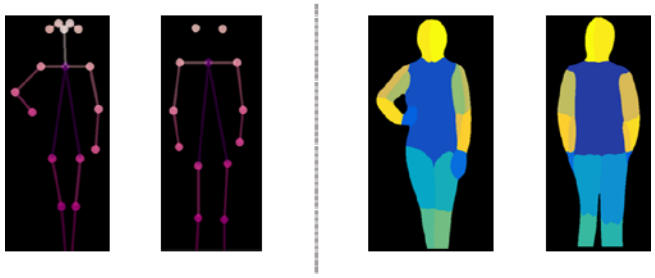
Ⅳ. 실 험

4.1 데이터

본 연구의 의복 가상 합성 딥러닝 모델에는 AIHub의 '패션상품 및 착용 영상' 이미지 데이터가 활용되었다(AIHub, 2020). 원천 데이터의 경우 의복 원본 이미지, 세그멘테이션, 포즈, 착용 원본 이미지, 세그멘테이션, 포즈 등의 데이터로 구성되어 있다. 본 제안 기법의 경우 Densepose 히트맵 형태의 신체 데이터가 모델에 입력되어야 하므로, DenseposeNet을 활용하여 필요한 이미지를 생성했다(Güler et al., 2018). Densepose 히트맵이란 컬러 이미지에서 신체의 영역을 더욱 세분화하는 방식으로, 기존의 스켈레톤 포즈(Skeleton Pose)에 비교적 정밀한 신체 구조를 표현할 수 있다. 본 연구의 모델 네트워크에서 추가적으로 필요한 이미지 레이블링 처리 및 대체 영역 처리를 통하여 <표 1>과 같이 총 5가지 종류의 이미지 데이터셋을 생성했다. 결과적으로 본 연구에서는 5개 분류 각 41,750개의 데이터를 학습 데이터로 적용했고, 검증 데이터로 각 4,721개를 사용하여 최종 232,355개의 이미지 데이터를 활용했다. 이미지 생성 모델의 특성상 많은 학습 데이터가 필요하여서 학습 데이터셋에 비교적 많은 비율로 데이터를 할당했다. 이미지 데이터의 모든 전처리가 끝난 후, 모델 입력을 위해서 640x640 크기로 변환했다. 이미지 크기의 경우 딥러닝 모델 학습에 필요한 컴퓨팅 자원을 고려하여 결정했다.

4.2 모델링

본 연구의 모델링은 먼저 의복 조건부 영역 생성기를 구축하는 것으로 시작했다. 조건부 영역 생성기의 경우 하이퍼 파라미터(hyper-parameter) 설정에서 배치 크기 8, 학습 스텝을 300,000으로 설정했다. GAN과 같은 생성 모델의 경우 학습 가중치의 수와 구조 복잡성이 일반적으로 높으므로 많은 스텝의 학습이 필요할 수 있다. 생성기 및 판별자의 학습률은 안정적인 학습을 위해서 0.0002로 설정했다. 모델 학습 시에는 MP 기법의 적용을 위해 손실 함수



<그림 5> Skeleton Pose(좌), Densepose Heatmap(우) 예시

<표 1> 모델 입력 데이터셋 예시

데이터셋 설명	예시 이미지
의복 제품 원본 이미지	
의복 제품 세그멘테이션	
착용자의 신체 Densepose 히트맵	
착용자 의복 대체 영역	
원본 착용자 이미지	

및 그래디언트 최적화 함수를 재정의했다. 그리고 모델에 데이터를 입력하기 위해서 <표 1>의 5가지 데이터와 이미지 레이블 정보를 로드했다. 딥러닝 모델 학습 절차에서는 인코더 구조를 통해서 학습된 의복 및 착용자 신체 정보를 <그림 3>과 같이 특성 융합 블록에 반영했고, 후처리를 통해서 <그림 4>와 같은 의복 조건부 영역 이미지를 생성했다. 최종 이미지 생성기 모델의 경우 하이퍼 파라미터 설정에서 배치 크기를 6으로, 학습 스텝은 150,000으로 설정했다. 최종 이미지 생성을 위한 디코더 네트워크 구조에서 안정성 보완을 위해서 의복 대체 영역 및 Densepose 히트맵 이미지 데이터가 입력되었다.

4.3 실험 결과

본 연구의 딥러닝 모델의 검증을 위해서 기존의 CP-VTON과의 AI 추론 결과를 비교 분석했다. 딥러닝 모델 검증 지표 분석 결과는 다음

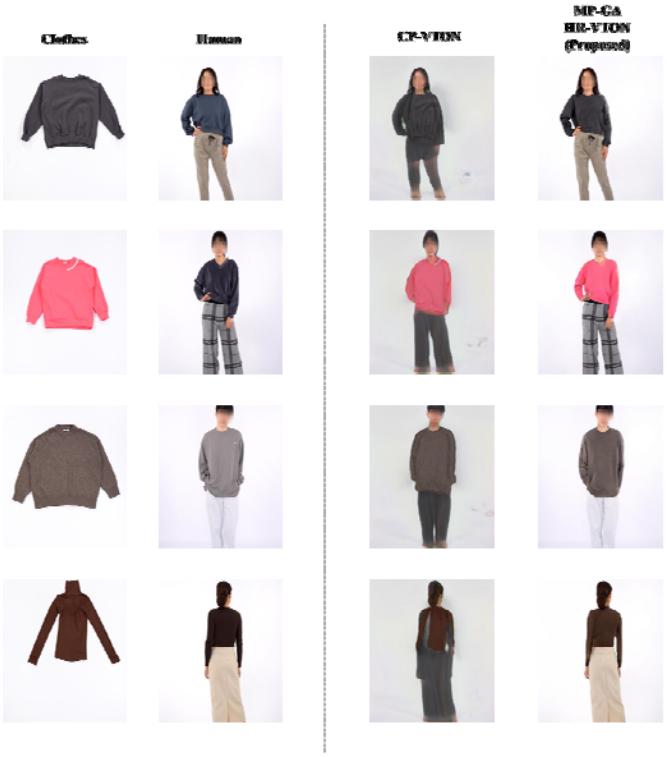
<그림 6> 의복 조건부 영역 생성 이미지 예시

과 같다. SSIM(Structural Similarity Index Map)의 경우 이미지 품질 측면의 특성(구조, 휘도, 대조)을 기준으로 실제 이미지와 생성 이미지의 유사성을 측정하는 것을 의미한다(Wang et al., 2003). 수치가 높을수록 실제 이미지와 생성 이미지의 구도적인 특성이 유사하다는 것을 의미한다. LPIPS(Learned Perceptual Image Patch Similarity)는 AlexNet 사전학습 모델을 기반으로 실제 이미지와 생성 이미지의 컨볼루션 연산 값을 유클리드 거리로 나타낸 지표이다(Zhang et al., 2018). 수치가 낮을수록 실제 이미지와 생성 이미지의 컨텍스트 정보가 유사하다는 것을 의미한다. FID(Frechet Inception Distance)란 사전학습 모델을 활용하여 학습된 현실의 이미지와 가상의 이미지 사이에서 컨볼루션 연산 값의 확률 분포를 비교하는 것을 의미한다(Zhang et al., 2018). 수치가 낮을수록 사람의 인지적인 관점에서 현실 세계의 이미지를 제대로 구현했다는 해석이 가능하다. KID (Kernel Inception Distance)는 실제 이미지와 가상의 이미지 픽셀값의 분포 거리를 분석하는 것을 의미한다. KID는 실제 및 가상 이미지 각각의 분포 차이와 실제 이미지와 가상 이미지 분포 차이를 고려하여 실제 이미지와 가상 이미지 사이의 차이, 즉 거리를 측정할 수 있다. 모든 지표에서 본 제안 기법의 성능이 우수한 것으로 검증 지표가 측정되었다. 이는 본 연구의 새로운 시도인 후면 이미지 학습 및 가중치 공유 방법론이 착용자 포즈 전/후면에 효과적으로 적용될 수 있는 것으로 보여진다.

<표 2> 의복 가상 합성 딥러닝 모델 검증 지표 분석 결과

Data Type	Metrics	CP-VTON	MP-GA HR-VITON (proposed)
Pair	SSIM	0.8385	0.9204
Pair	LPIPS	0.2133	0.0642
Unpair	FID	74.5421	11.8463
Unpair	KID	0.064	0.006

본 연구의 AI 모델의 추론 결과는 <그림 5>와 같다. 의복과 착용자의 포즈가 MP-GA HR-VITON 모델에 입력되었을 때 의복의 질감, 색상, 굴곡 등 자연스러운 합성 결과를 도출했다. 특히 <그림 5>의 네 번째 예시에서 후면 의복 및 포즈를 실제 사진과 같이 합성할 수 있다는 것을 증명했다. 본 연구는 기존 연구와 같은 정형화된 의복 고스트컷 이미지가 아니라 가이드 라인 없이 구비된 의복 제품 이미지를 활용했다(Han et al., 2018; Lyu et al., 2021). 네 번째 후면 이미지 합성 예시의 경우, 의복



<그림 7> 본 연구의 AI 모델 기반 의류 가상 합성 예시

제품 이미지에 상당한 왜곡이 있음에도 불구하고 AI 모델을 통해서 자연스러운 합성이 가능했다. 본 연구의 AI 모델은 단색 의복 이미지에서 높은 수준의 합성 품질을 달성했다. CP-VTON의 경우 후면 이미지가 모델에 입력되면서 합성 영역을 대부분 포착하지 못한다는 결과를 도출했다. 이는 본 연구의 전면 및 후면 모델 가중치 공유 방식이 의복 조건부 영역 생성기에 효과적으로 적용되었다는 것을 의미한다. 또한, 본 연구는 HR-VITON 방법론을 기반하기 때문에 기존 연구와 비교적 고해상도인 640x640 이미지를 효율적으로 학습할 수 있었다. 모델 학습 컴퓨팅 자원 및 시간에 따라서 초고화질 이미지를 학습하는 것도 가능할 것으로 판단된다.

Ⅴ. 결 론

본 연구에서는 기존의 VTON 연구의 한계점을 보완하기 위해 가중치 공유 및 MP-GA 기법을 활용하여 전면 및 후면 의복 착용 이미지를 효율적으로 학습할 수 있는 방법론을 제시했다. 본 연구의 학습 목적에 따라서 HR-VITON 모델을 재설계했으며, 모델 학습에 필요한 AIHub의 '패션 상품 및 착용 영상' 이미지 데이터 232,355건을 원천 데이터로부터 생성했다. 해당 이미지 데이터는 의복 제품 원본 이미지, 의

<그림 8> Bad Sample 예시

복 제품 세그멘테이션, 착용자의 신체 Densepose 히트맵, 착용자 의복 대체 영역, 원본 착용자 이미지 데이터셋을 포함한다. AI 모델의 정밀한 의복 이미지 합성을 위해서 많은 이미지 데이터셋 및 처리 과정이 필요했다. 모델링에서는 의복 조건부 영역 생성기에서 배치 크기 8, 학습 스텝을 300,000으로 설정하고, 최종 이미지 생성기 모델의 경우 배치 크기를 6으로, 학습 스텝은 150,000으로 설정했다. MP-GA 등의 학습 최적화 기법이 적용되었음에도 모델링 학습 실험에서 평균적으로 일주일의 시간이 필요했다. <표 2>와 <그림 5>처럼, 본 연구의 제안 모델이 기존 연구의 방법론에 비교적 우수한 검증 지표를 달성했으며, 다양한 포즈, 색상, 질감, 전면 및 후면 이미지를 합성할 수 있음을 나타냈다.

본 연구의 시사점은 다음과 같다. 첫 번째, 해외에서 활발하게 연구되고 있는 딥러닝 기반 의류 자동 합성 분야의 최신 기술을 활용하여 국내 데이터를 대상으로 실험 결과를 실증했다는 것이다. 기존 국내 연구의 경우, VITON의 초기 모델을 활용하여 합성 품질이 저조할 수 있지만, 본 연구는 고해상도 이미지 학습을 효과적으로 수행할 수 있는 HR-VITON 모델을 응용했다(AIHub, 2020, Lyu et al., 2021). 딥러닝 비전 분야의 최근 동향으로 트랜스포머(transformer) 기법을 응용하는데, 향후 연구에서는 트랜스포머 기반의 의류 합성 모델을 연구하고자 한다(Vaswani et al., 2017; Dosovitskiy et al., 2020; Wang et al., 2022). 본 연구의 딥러닝 모델을 활용하여 추론 시 단색 옷은 자연스러운 합성이 가능하지만, <그림 6>과 같이 복잡한 문양, 로고 등의 요소가 있을 때 합성 시 부자연스러운 패턴이 발생했다. 이를 트랜스포머를 기반으로 고도화한다면 이와 같은 문제점도 해결할 수 있을 것으로 예상된다.

두 번째, 본 연구의 딥러닝 모델은 전면 및 후면 학습 모델을 독립적으로 설계 후 가중치를 공유하는 방식을 활용했다. 기존 연구에서는 대부분 전면 이미지 데이터셋을 활용했는데, 본 연구는 AIHub의 후면 이미지 데이터셋을 실험적으로 모델 학습에 활용했다(Han et al., 2018; Lyu et al., 2021). 그러나 전면 및 후면 이미지의 경우 의복과 착용자의 신체 구조가 비교적 정형화되어 있지만, 측면 이미지의 경우 구도상 신체 구조가 매우 불규칙하여서 합성에 어려움이 있었다. 측면 의복 이미지를 자연스럽게 합성하기 위해서는 새로운 접근 방식이 필요하다. 세 번째, 본 연구는 기존 연구와 다르게 평면상

의 의복 제품 이미지를 합성했다(Han et al., 2018; Lyu et al., 2021). 고스트컷 의복 이미지를 활용하면 사람의 신체 구조를 효과적으로 반영할 수 있지만, 가상 합성 서비스의 측면에서는 비효율적인 상황을 발생시킬 수 있다. 비즈니스적인 관점에서 VTON 모델은 다른 AI 및 데이터 처리 없이 의복 제품 및 착용자의 사진만으로 합성하는 기술이 가장 큰 효용을 가지고 있다고 볼 수 있다. 마지막으로 본 연구는 MP-GA 기법을 통해서 딥러닝 모델의 학습과 비용을 최적화했다. 해당 기법을 적용하기 전에는 컴퓨팅 자원의 메모리 문제로 배치 크기를 2로 설정하여 수행하려고 했으나, MP-GA 적용 후 배치 크기를 6~8 수준으로 설정하는 것이 가능했다. 이에 따라서 모델 학습 예상 시간을 추정했을 때, MP-GA를 적용하기 전 학습 예상 시간이 285시간, 적용 후 160시간으로 나타났다.

<표 2> 의복 가상 합성 딥러닝 모델 검증 지표 분석 결과

개발 PC 사양	CPU: Intel(R) Xeon(R) Gold 6132 16 Core 2.60GHz, RAM: 120GB, GPU: V100 VRAM 32GB
Image Generator ETA	가용 batch_size 6, ETA 160hours
MP-GA Image Generator ETA	가용 batch_size 2, ETA 285hours

VTON 모델이 많은 데이터를 오랜 시간 학습하는 과정이 필요한 만큼, 향후 연구에서는 추가적인 모델 학습 최적화, 분산 및 병렬 학습 기법을 활용하고자 한다. 본 연구를 통해서 국내 의복 이미지 자동 합성 연구 분야와 패션 커머스, 최종적으로 국내 섬유 산업에서의 빅데이터 플랫폼의 활성화를 기대한다.

참고문헌

매일신문, “다이텍연구원, ‘비대면 섬유소재 마케팅 플랫폼’ 만든다”, 신중언 기고, 2020.12.29.

신은경, 김은미, 홍태호, “Som과 Lstm을 활용한 지역기반의 부동산 가격 예측”, 정보시스템연구, 제30권, 2호, 2021, pp. 147-163.

원종관, 홍태호, 배경일, “신용 데이터의 이미지 변환을 활용한 합성곱 신경망과 설명가능한 인공지능(XAI)을 이용한 개인신용평가”, 정보시스템연구, 제30권, 4호, 2021, pp. 203-226.

조보근, 박경배, 하성호, “기계학습 알고리즘을 활용한 지역별 아파트 실거래가격지수 예측모델 비교: Lime 해석력 검증”, 정보시스템연구, 제 29, 3권, 2020, pp. 119-144.

AIHub., “패션상품 및 착용 영상,” https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=78, 2020.

Ayush., C., “Project Clothes Swap,” *2020 Adobe Summit Sneaks Session*, https://business.adobe.com/summit/2020/clothes-swap-summit-sneak.html, 2020.

Cao, Z., Tomas, S., Shih-En, W., and Yaser, Sheikh., “Realtime Multi-Person 2d Pose Estimation Using Part Affinity

Fields," *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Dosovitskiy, A., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., and Mostafa, D., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv Preprint ArXiv:2010.11929,* 2020.

Gong, K., Xiaodan, L., Dongyu, Z., Xiaohui, S., and Liang, L., "Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing," *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Güler, R, A., Natalia, N., and Iasonas, K., "Densepose: Dense Human Pose Estimation in the Wild," *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Han, X., Zuxuan, W., Zhe, W., Ruichi, Y., and Larry, S, D., "Viton: An Image-Based Virtual Try-on Network," *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Lee, S., Gyojung G., Sunghyun P., Seunghwan C., and Jaegul C., "High-Resolution Virtual Try-on with Misalignment and Occlusion-Handled Conditions," *Paper Presented at the European Conference on Computer Vision*, Github: https://github.com/sangyun884/hr-viton, 2022.

Lin, C., Zhao L., Sheng, Z., Shichang, H., Jialun, Z., Linhao, L., Jiarun, Z., Longtao, H., and Yuan, H., "Rmgn: A Regional Mask Guided Network for Parser-Free Virtual Try-On," A*rXiv Preprint ArXiv:2204.11258*, 2022.

Lyu, Q., Qiu-Feng, W., and Kaizhu, H., "High-Resolution Virtual Try-on Network with Coarse-to-Fine Strategy," *Paper Presented at the Journal of Physics: Conference Series*, 2021.

Micikevicius, P., Sharan, N., Jonah, Alben., Gregory, D., Erich, E., David, G., and Boris, Ginsburg., "Mixed Precision Training," *ArXiv Preprint ArXiv: 1710.03740,* 2017.

Ruder, S., "An Overview of Gradient Descent Optimization Algorithms," A*rXiv Preprint ArXiv:1609.04747*, 2016.

Vaswani, A., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N, G., Łukasz, K., and Illia P., "Attention Is All You Need," *Paper Presented at the Advances in Neural Information Processing Systems*, 2017.

Wang, B., Huabin, Z., Xiaodan, L., Yimin, C., Liang, L., and Meng, Y., "Toward Characteristic-Preserving Image-Based Virtual Try-on Network," *Paper*

Presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018.

Wang, W., Hangbo, B., Li, D., Johan, B., Zhiliang, P., Qiang, L., and Kriti, A., "Image as a Foreign Language: Beit Pretraining for All Vision and Vision-Language Tasks," A*rXiv Preprint ArXiv:2208.10442*, 2022

Wang, Z., Eero, P, S., and Alan, C, B., "Multiscale Structural Similarity for Image Quality Assessment," *Paper Presented at the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003.

Yang, H., Ruimao, Z., Xiaobao, G., Wei, L., Wangmeng, Z., and Ping, L., "Towards Photo-Realistic Virtual Try-on by Adaptively Generating-Preserving Image Content," *Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Zhang, H., Feng, L., Shilong, L., Lei, Z., Hang, S., Jun, Z., Lionel, M, N., and Heung-Yeung, S., "Dino: Detr with Improved Denoising Anchor Boxes for End-to-End Object Detection," *ArXiv Preprint ArXiv:2203.03605*, 2022.

Zhang, R., Phillip, I., Alexei, A, E., Eli, S., and Oliver, Wang., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

이 현 상 (Lee, Hyun Sang)

경북대학교 경영학사와 석사학위를 취득하였다. 현재 경북대학교 박사과정을 수료했으며, 주요 관심분야는 경영정보, 딥러닝, 자연어 처리, 시계열 예측 등이다.

오 세 환 (Oh, Se Hwan)

현재 경북대학교 경영학부에서 부교수로 재직 중이다. 서울대학교 경제학부(학사)를 졸업했으며 카네기멜론대에서 e-비즈니스 석사, 서울대학교에서 경영학(경영정보) 박사학위를 받았다. 주요 연구 관심분야는 전자상거래, 국제무역에서의 IT 활용 등이다.

하 성 호 (Ha, Sung Ho)

연세대학교 경영학사와 한국과학기술원 석사와 박사학위를 취득하였다. 현재 경북대학교 경영학부 교수로 재직하고 있으며, 주요 관심 분야는 경영 분야에 적용되는 데이터 사이언스, 기계학습, 딥러닝 등이다.

<Abstract>

Virtual Fitting System Using Deep Learning Methodology: HR-VITON Based on Weight Sharing, Mixed Precison & Gradient Accumulation

Lee, Hyun Sang · Oh, Se Hwan · Ha, Sung Ho

Purpose

The purpose of this study is to develop a virtual try-on deep learning model that can efficiently learn front and back clothes images. It is expected that the application of virtual try-on clothing service in the fashion and textile industry field will be vitalization.

Design/methodology/approach

The data used in this study used 232,355 clothes and product images. The image data input to the model is divided into 5 categories: original clothing image and wearer image, clothing segmentation, wearer's body Densepose heatmap, wearer's clothing-agnosting. We advanced the HR-VITON model in the way of Mixed-Precison, Gradient Accumulation, and sharing model weights.

Findings

As a result of this study, we demonstrated that the weight-shared MP-GA HR-VITON model can efficiently learn front and back fashion images. As a result, this proposed model quantitatively improves the quality of the generated image compared to the existing technique, and natural fitting is possible in both front and back images. SSIM was 0.8385 and 0.9204 in CP-VTON and the proposed model, LPIPS 0.2133 and 0.0642, FID 74.5421 and 11.8463, and KID 0.064 and 0.006. Using the deep learning model of this study, it is possible to naturally fit one color clothes, but when there are complex pictures and logos as shown in <Figure 6>, an unnatural pattern occurred in the generated image. If it is advanced based on the transformer, this problem may also be improved.

Keyword: Virtual Try-on Service, Deep learning, HR-VITON, Weight Sharing, Mixed Precision,

Gradient Accumulation

* 이 논문은 2022년 11월 4일 접수, 2022년 11월 29일 1차 심사, 2022년 12월 21일 게재 확정되었습니다.