PDR 보고서

주제	FitPlus
팀명	유진경은호
작성자	진경은

1. 개요

(1) 서비스 설명

FitPlus는 온라인 쇼핑몰에서의 사이즈 불일치로 인한 반품률 감소와 소비자 만족도 향상을 위한 플랫폼입니다. 사용자의 신체 정보를 바탕으로 적절한 사이즈가 자동 추천되며, FitPlus는 딥러닝 기반 이미지 합성 알고리즘을 활용해 자연스럽게 착용할 수 있는 가상 피팅 결과를 제공합니다. 이 플랫폼은 사용자가 직접 제품 이미지를 캡처하거나 온라인에서 선택한 의류에 대해 가상 피팅을 실시하여 반품률을 줄이고 소비자의 구매 결정을 더 쉽게 내릴 수 있습니다. FitPlus의 모든 데이터 처리 과정이 스마트폰 내에서 이루어지므로 개인정보 보안과 데이터 보안 측면에서도 높은 신뢰성을 제공합니다.

(2) 서비스의 필요성

FitPlus는 온라인 쇼핑몰에서 발생하는 높은 반품률 문제를 해결하기 위해 설계된 On-Device 가상 피팅 서비스입니다. 특히, 패션 의류의 반품 사유 중 약 54%가 사이즈 불일치와 관련되어 있으며, 이는 e커머스 업계의 주요 과제로 꼽히고 있습니다. 기존 오프라인 매장의 평균 반품률(10%)보다 온라인 쇼핑몰의 반품률(30%)이 세 배 이상 높은 이유도 소비자가 실제로 착용해볼 수 없기 때문입니다[1][2]. 또한, 사이즈 불일치 문제로 인해 '브래킷팅(Bracketing)'이라는 구매 현상도 발생하고 있습니다. 이는 여러 사이즈를 동시에 구매한 후, 맞지 않는 제품을 반품하는 방식으로, 온라인 쇼핑몰의 물류 부담과 반품률을 더욱 가중시키고 있습니다.

· 사이즈 문제로 인한 반품률 감소

FitPlus는 사용자가 선택한 의류를 자신의 사진에 합성하여 실제 착용한 모습과 유사한 경험을 제공합니다. 이를 통해 소비자는 구매 전에 옷의 핏과 스타일을 미리 확인할 수 있어, 사이즈 불일치로 인한 반품을 효과적으로 줄일 수 있습니다.

· 소비자 만족도 및 구매 결정 지원

온라인 쇼핑은 상품을 직접 확인하지 못한다는 점에서 소비자에게 불확실성을 제공합니다. FitPlus는 이러한 한계를 극복하기 위해 정확하고 자연스러운 가상 피팅 경험을 제공하여 소비자가 구매 결정을 더 쉽게 내릴 수 있도록 돕습니다. 이는 소비자 만족도를 높이고 재구매율 향상에도 기여할 수 있습니다.

· 데이터 보안과 개인 정보 보호

FitPlus는 모든 이미지 처리 과정을 사용자의 스마트폰 내에서 수행하기 때문에, 사용자의 승인 없이 외부 서버로 데이터를 전송하지 않아 개인 정보 보호와 데이터 보안 측면에서도 높은 신뢰성을 제공합니다. 이는 디지털 환경에서 점점 더 중요해지는 보안 요구를 충족시키며, 소비자가 안심하고 서비스를 이용할 수 있도록 합니다.

[1] 한국경제, perfitt 관련. https://www.hankyung.com/article/2024080531631
[2] 물류신문, 반품 관련. https://www.klnews.co.kr/news/articleView.html?idxno=312851
[3] perfitt, e커머스 버츄얼 피팅 관련. https://www.perfitt.io/#info

2. 사용자 가상 시나리오 및 요구사항 정리

(1) 사용자 가상 시나리오

	사용자 프로필	사용 상황	요구 기능
1	김하늘 (27세, SNS 마케터)	인스타그램 룩북 콘텐츠 촬영 전, 여러 코디 조합을 시뮬레이션	✔ SNS 공유 기능 ✔ 착장 이미지 저장
2	정민석 (30세, 회사원)	출근용 셔츠 3종을 한 번에 비교해보고 가장 어울리는 스타일 선택	✔ 여러 의류 비교 기능 ✔ 실시간 착장 비교
3	이세영 (26세, 패션 유튜버)	브랜드별 청바지 S, M, L 사이즈를 실측치 기준으로 비교	✔ 실측 기반 사이즈 비교 ✔ 사이즈별 핏 시뮬레이션
4	박지훈 (33세, 신혼부부)	커플룩 코디를 함께 보고 마음에 드는 룩만 따로 저장	✔ 마음에 드는 착장 저장 ✔ 커플 피팅 모드
5	송은지 (29세, 온라인 쇼핑 애호가)	지금까지 구매한 브랜드와 핏을 참고해 새 상품 추천받기	✔ 구매 이력 기반 추천 ✔ 이전 착장과 비교 기능
6	최윤아 (24세, 대학생)	동아리 발표회 의상 후보 5벌 중 친구들과 함께 투표	✔ 피팅 룸 공유 기능 ✔ 착장 투표 기능
7	장혁 (36세, 육아휴직 아빠)	아이와 함께하는 가족 나들이 옷을 함께 비교하고 저장	✔ 가족 피팅 지원 ✔ 비교 + 저장 기능
8	백수진 (41세, 초등학교 교사)	강의용 셋업 2벌을 비교하고, 지난 학기 구매와 핏 비교	✔ 분석 ✔ 비교 기능
9	임태호 (38세, 운동 마니아)	운동복 구매 시 체형 맞춤 사이즈와 디자인 비교	✔ 실측 기반 사이즈 피팅 ✔ 활동성 중심 스타일 비교
10	고지민 (32세, 프리랜서 디자이너)	매일 입는 기본 아이템을 여러 색상·핏으로 비교	✔ 컬러별 비교 기능 ✔ 마음에 드는 코디 저장

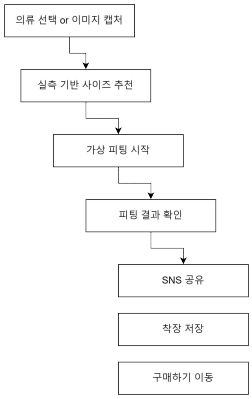
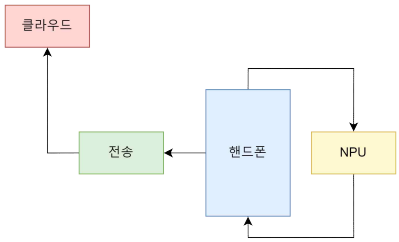
(2) 사용자 요구사항

기능 분류	고객 요구사항 내용
착장 기능	착장 이미지 및 코디를 SNS에 바로 공유할 수 있어야 함
비교 기능	여러 옷을 동시에 피팅하고 비교할 수 있는 인터페이스
실측치 기반 비교	사용자의 신체 치수 기반으로 다양한 사이즈를 정확히 시뮬레이션
착장 기능	마음에 드는 착장을 저장하고 나중에 다시 확인 가능
투표 및 공유	친구, 연인과 피팅 룸 공유 및 의견 교환 가능
가족, 연인 단위 지원	가족 구성원별 피팅 비교 및 커플룩, 가족룩 추천

(3) 기능 우선순위 매트릭스

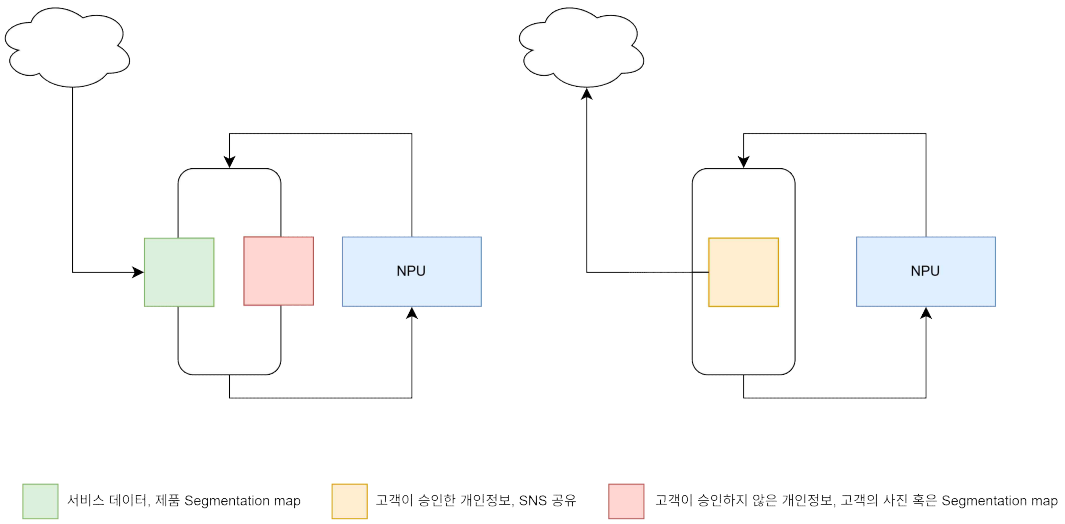
기능명	사용자 가치	개발 난이도	우선순위 분류	비고
착장 기능	매우 높음	높음	1	핵심 기능
실측 기반 사이즈 비교	매우 높음	높음	1	핵심 기능
다중 사용자 착장 기능	매우 높음	매우 높음	2	부가적인 기능

(4) 실행 흐름 및 구조도

실행 흐름도	기본 구조도
	
사용자는 온라인 쇼핑몰에서 마음에 드는 옷을 선택하거나, 직접 제품 이미지를 캡처하여 피팅을 시작할 수 있습니다. 사용자가 입력한 키, 몸무게 등의 신체 정보를 바탕으로 적절한 사이즈가 자동으로 추천됩니다. 이 과정은 반품률을 줄이는 핵심 기능입니다. 추천된 사이즈를 바탕으로, 사용자의 사진에 선택한 의류가 자연스럽게 합성되어 착용한 모습이 시뮬레이션 됩니다. 사용자는 가상 피팅 결과를 통해 스타일, 핏, 색상 등을 실제 착용한 듯한 이미지로 확인할 수 있습니다. 피팅 결과를 바탕으로 사용자는 다양한 행동을 선택할 수 있습니다. 착장 이미지를 인스타그램, 카카오톡 등 SNS에 공유하여 피드백을 받을 수 있습니다. 마음에 드는 착장은 저장해두고 나중에 다시 확인하거나 비교할 수 있습니다. 최종적으로 만족스러운 스타일을 결정하면 쇼핑몰로 바로 이동하여 구매할 수 있습니다.	사용자는 핸드폰 내에서 가상 피팅을 위한 모든 과정을 실행합니다. 고성능 스마트폰의 NPU를 활용해, 딥러닝 기반 이미지 합성 알고리즘을 빠르게 실행할 수 있습니다. 이 과정을 통해 사용자의 신체 이미지와 의류 이미지가 결합되어 자연스러운 착장 이미지가 생성됩니다. FitPlus는 모든 주요 연산을 사용자의 디바이스 내에서 수행합니다. 외부 서버나 클라우드로 원본 이미지가 전송되지 않기 때문에, 민감한 신체 이미지나 얼굴 정보가 의도치 않게 노출될 위험이 없습니다. 이는 최근 강화되고 있는 개인정보 보호 요구사항과도 부합하며, 사용자의 신뢰를 높이는 핵심 요소입니다. 최신 스마트폰에 탑재된 NPU는 AI 연산 최적화에 특화되어 있으며, 딥러닝 기반의 딥러닝 모델을 실시간으로 구동할 수 있는 성능을 갖추고 있습니다. 기존 클라우드 기반 처리 방식보다 빠르고 안정적이며, 네트워크 연결 없이도 오프라인 환경에서도 동작이 가능합니다. 사용자가 SNS 공유나 데이터 백업 등의 기능을 선택한 경우에만, 결과 이미지 또는 가공된 데이터가 클라우드로 전송됩니다. 이때에도 민감한 원본 정보는 포함되지 않으며, 최소한의 데이터만 전송되도록 설계되어 있어 보안성과 효율성을 모두 충족합니다.

3. 데이터 처리 방식 관련 세부 구조도

세부 구조도



해당 내용은 기본 구조도와 이어지는 내용입니다. 개인정보 처리 가이드에 따라 데이터를 3가지로 분류했습니다. 먼저 서비스 데이터는 클라우드 서버에서 연산 및 처리되어 제공되는 데이터로 이 경우 사용자가 입력한 제품의 사진을 딥러닝에 적합하게 전처리한 특징벡터를 의미합니다. 고객이 승인하지 않은 개인정보는 온디바이스 환경에서만 연산되는 데이터로 사용자의 사진과 타겟 지점으로 사용될 Segmentation map을 의미합니다. 고객이 승한 개인정보는 고객의 신체 정보를 의미하며 고객의 키, 몸무게, 어깨선의 길이가 이에 해당하고 고객의 사이즈 추천 시스템에 사용됩니다. 데이터 처리 방식의 차이를 사용자가 명확히 인지할 수 있도록 제품에 대한 후처리 과정은 클라우드 서버에서 처리하여 제공하고 전체적인 이미지 합성 과정에 타겟 지점으로 활용할 사용자 본인의 Segmentation map은 사용자가 직접 그릴 수 있도록 구성할 계획입니다.

(1) 데이터 수집

• 의류 사이즈 표기법

KS 의류 치수 규격을 사용합니다. 다만 달라붙는 옷과 달라붙지 않는 옷 등 옷의 종류에 따라 그 규격이 달라집니다. 또한 국가별 표기법에도 차이가 있습니다. 따라서 실측치(cm)를 기준으로 직접 치수를 측정한 무신사 데이터를 크롤링해 사용합니다.

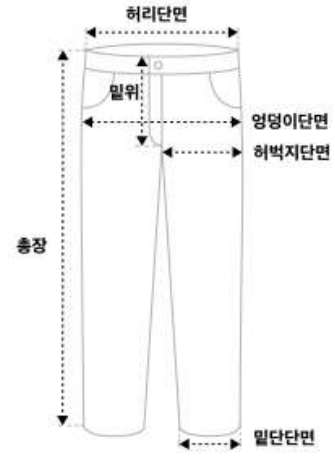
• 무신사 홈페이지 - 사이즈 관련 데이터

무신사 홈페이지는 자바스크립트를 통해 동적으로 html 문서를 생성합니다. 따라서 selenium을 사용해 크롤링합니다. 우선 수집 대상인 의류 종류는 상의, 아우터, 바지, 스커트이고 추가 수집 대상인 의류 종류는 원피스, 스포츠(레깅스, 민소매 등), 악세서리입니다.

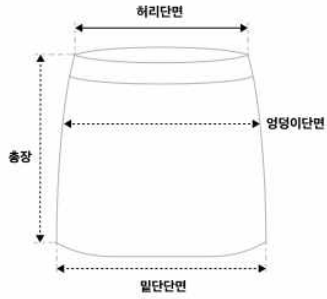
<table>
<tr><th>아우터</th><th>상의(긴팔)</th><th>상의(반팔)</th></tr>
<tr>
<td> 

사이즈 실측 안내 점퍼 사이즈 측정법</td>
<td> 

사이즈 실측 안내 긴소매티셔츠 사이즈 측정법</td>
<td> 

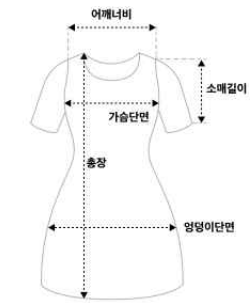
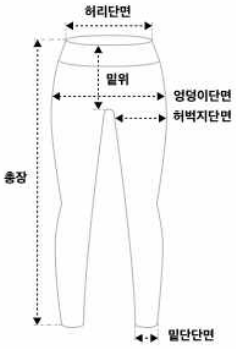
사이즈 실측 안내 반소매티셔츠 사이즈 측정법</td>
</tr>
<tr><th>하의(긴바지)</th><th>하의(반바지)</th><th>하의(치마)</th></tr>
<tr>
<td> 

사이즈 실측 안내 바지 사이즈 측정법</td>
<td> 

사이즈 실측 안내 반바지 사이즈 측정법</td>
<td> 

사이즈 실측 안내 스커트 사이즈 측정법</td>
</tr>
</table>

[표1] 주요 상품 실측치 규격

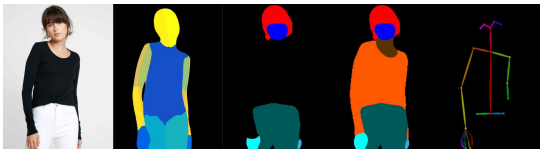
원피스	가방	민소매	레깅스
 사이즈 실측 안내 원피스 사이즈 측정법	 사이즈 실측 안내 토트/핸드백 사이즈 측정법	 사이즈 실측 안내 민소매 사이즈 측정법	 사이즈 실측 안내 레깅스 사이즈 측정법

[표2] 추가 상품 실측치 규격

• 공공 데이터 (AI hub) - 의류 관련 데이터

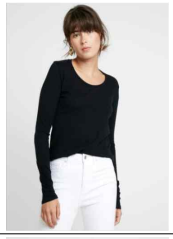
의류 통합 데이터 관련	디자인 패턴 관련
	
https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71501	https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71445

• VITON-HD – virtual try on 관련 데이터

모델 이미지	착장 의상
	
https://www.kaggle.com/datasets/marquis03/high-resolution-viton-zalando-dataset?select=train_pairs.txt	https://www.kaggle.com/datasets/marquis03/high-resolution-viton-zalando-dataset?select=train_pairs.txt

관련 링크 : https://github.com/shadow2496/VITON-HD

(2) 오픈 소스 benchmark

사용한 모델	모델 이미지	의상	결과물
HR-VITON https://github.com/sangyun884/HR-VITON 1024×768 A6000, 0.25초 1024×768 4090, 0.125초			
StableVITON[2] https://github.com/rlawjdghek/StableVITON 1024×768 A6000, 40초 4090, 35초 512×384 4090, 5초			
IDM-VTON[3] https://github.com/yisol/IDM-VTON 1024×768 A6000, 30초 1024×768 4090, 5초			
CatVITON[4] https://github.com/Zheng-Chong/CatVTON/ 1024×768 4090, 11초 512×384 4090, 1초			
Leffa[5] https://github.com/franciszzj/Leffa 1024×768 4090, 6초			

CatVTON 논문 기준(A100 한장)

Methods	GFLOPs				Inference Time(s)		Memory Usage	
	E_{text}	E_{image}	ReferenceNet	UNet	512×384	1024×768	512×384	1024×768
OOTDiffusion (Xu et al., 2024)	13.08	155.62	509.12	547.34	4.76	36.23	6854 M	8892 M
IDM-VTON (Choi et al., 2024)	110.04	155.62	1340.15	1163.98	12.96	17.32	17112 M	18916 M
StableVTON (Kim et al., 2023)	-	155.62	173.80	545.27	12.17	36.10	9828 M	14176 M
CatVTON(Ours)	-	-	-	973.59	2.58	9.25	3276 M	5940 M

Methods	Params (M)							Memory	Conditions	
	VAE	*UNet*	$UNet_{ref}$	E_{text}	E_{image}	Total	Trainable	Usage(G)	Pose	Text
OOTDiffusion (Xu et al., 2024)	83.61	859.53	859.52	85.06	303.70	2191.42	1719.05	10.20	-	✓
IDM-VTON (Choi et al., 2024)	83.61	2567.39	2567.39	716.38	303.70	6238.47	2871.09	26.04	✓	✓
StableVTON (Kim et al., 2023)	83.61	859.41	361.25	-	303.70	1607.97	500.73	7.87	✓	-
StableGarment (Wang et al., 2024c)	83.61	859.53	1220.77	85.06	-	2248.97	1253.49	11.60	✓	✓
MV-VTON (Wang et al., 2024a)	83.61	859.53	361.25	-	316.32	1620.71	884.66	7.92	✓	-
CatVTON (Ours)	83.61	**815.45**	-	-	-	**899.06**	**49.57**	**4.00**	-	-

(3) HR-VITON

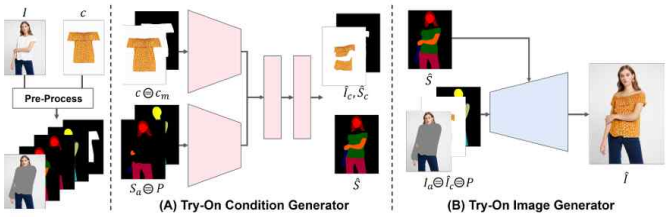


Fig. 2: Overview of the proposed framework (HR-VITON).

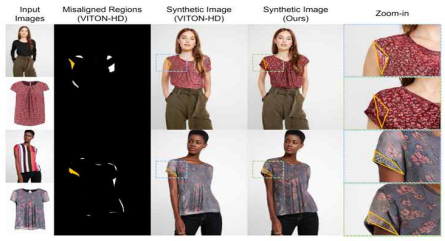
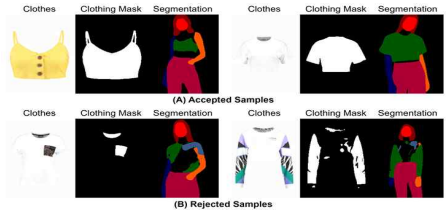
핵심 아이디어

1) Try-on Condition Generator

기존과 달리 의류 워핑 + 세그멘테이션 맵 생성을 하나의 네트워크에서 동시 수행. 두 작업의 출력을 Feature Fusion Block으로 워핑된 옷 위치와 세그멘테이션 경계가 어긋나 빈 공간이나 겹침이 발생하는 현상인 misalignment와 팔이나 머리 등과 같이 신체가 옷을 가릴 때 생기는 변형(Pixel-squeezing) 완화하는 효과.

2) Discriminator Rejection

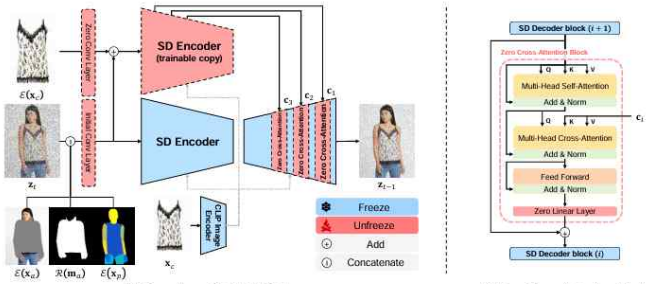
최종 생성 단계 이전에 Segmentation map을 예측하여 품질이 떨어지는 Segmentation map을 판별기(Discriminator)로 걸러 합성 단계 생략.

 Fig. 6: Synthesis results and corresponding misaligned regions indicated by yellow colored areas. VITON-HD suffers from the artifacts caused by misalignment.	misalignment
 Fig. 7: Effects of the body part occlusion handling. The green colored areas indicate the pixel-squeezing artifacts.	pixel-squeezing 현상
 Fig. 8: Examples of accepted (A) and rejected (B) segmentation maps by discriminator rejection, corresponding input clothes and clothing masks.	정상(a), 비정상(b) 세그맨테이션 결과

장점 및 단점

GAN 기반임. 매우 빠른 처리 속도(A6000 기준 장당 0.25초), Diffusion 기반에 비해 사실적 묘사가 부족함.

(4) StableVITON



(a) Overview of StableVITON (b) Zero Cross-Attention Block

핵심 아이디어

1) 인페인팅 문제

Stable Diffusion 백본을 그대로 사용. 가상피팅을 인페이팅 문제로서 접근함.

2) Zero Cross-Attention

인코더에서 추출한 의류 토큰을 Key, Value로 사용 사람(Agnostic map 등)을 Query로 사용하여 워핑 네트워크 없이 옮겨 붙일 위치를 토큰 단위로 학습함. Zero의 의미는 Key를 만드는 weight와 value를 만드는 weight를 0으로 초기화하여 모델 미세조정 과정에서 원래 지식을 보존하며 안정적인 학습이 가능하게끔 하기 위함.

3) Attention Total-Variation Loss

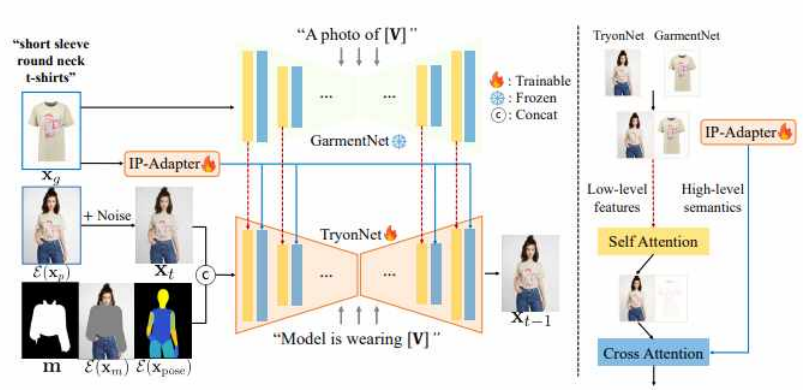
Zero Cross-Attention으로 의류 토큰(Key/Value)과 사람 토큰(Query)을 매칭을 수행. 하지만 학습 초기에 attention score가 군데군데 흩어져 옷의 색감이나 세밀한 부분이 오류가 발생. 따라서 attention map이 덩어리를 이루도록 부드럽게 만들어 주는 역할을 함.

장점 및 단점

Diffusion 기반. 처리 속도(A6000, 장당 1초), GAN 기반에 비해 높은 사실감. 별도의 워핑 과정이 없으므로 misalignment 문제가 자연스럽게 완화. Occlusion에 약함. 입력 해상도가 낮으면 품질 저하.

 Figure 2. Visualization of the semantic correspondence learned by our StableVITON. We overlay the attention map for the clothing regions onto the generated images for visualization.	Zero Cross-Attention을 활용한 정합과정
 Figure 4. Visualization of attention map from a zero cross-attention block of 32 resolution.	Attention Total-Variation Loss의 효과

(5) IDM-VTON



핵심 아이디어

1) 다양한 정보의 합성

TryonNet(SDXL Unet)을 중심으로, 의류 정보를 두 갈래로 주입. IP-Adapter는 CLIP 기반 이미지 임베더로서 의류의 고수준 정보(종류, 색 등) 제공함. GarmentNet은 TryonNet과 같은 SDXL Unet이고 의류의 저수준 정보(질감, 로고, 무늬 등)을 추출함. 의류마다 상세한 텍스트 캡션을 함께 입력해 사용.

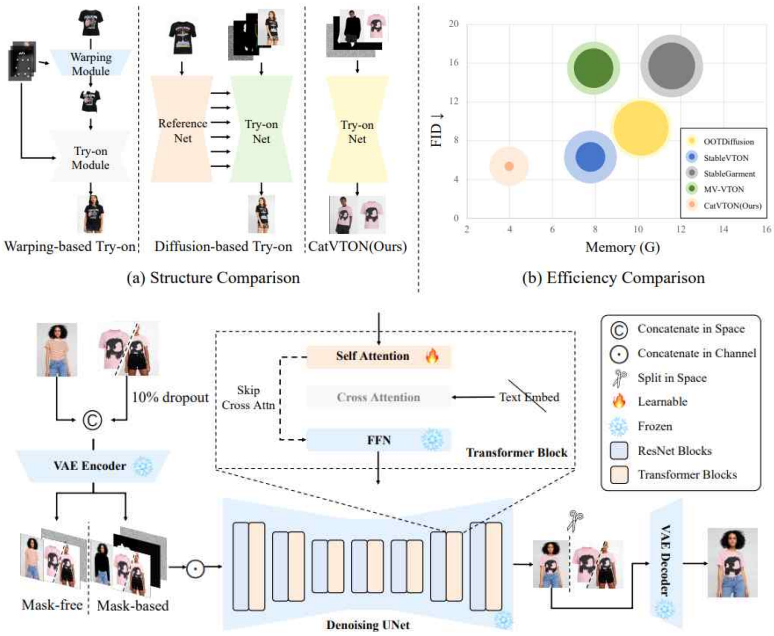
2) 효율적인 학습

옷 사람 한 쌍 이미지로 TryonNet 디코더의 attention 부분만 재학습함. 배포 시 약 100 step 파인튜닝만으로 실제 촬영 옷에도 적용 가능.

장점 단점 한계점

저수준 고수준 이중 조건을 활용하여 SOTA를 기록함. 미세조정 비용이 비교적 적음.
모델의 구조가 복잡하고 옷으로 가려졌던 부분이 복원되지 않음. 의류를 자세히 묘사한 캡션만 사용함. 주로 상의 위주로 연구가 진행됨.

(6) CatVTON



핵심 아이디어

1) 경량 구조

하나의 VAE Encoder-UNet을 통해 사람 이미지와 의류 이미지를 concat하여 공동처리하는 방식을 사용함. Denosing UNet 부분에 텍스트Encoder, reference UNet(GarmentNet 등) CLIP/DINOv2등 추가 모듈 제거.

2) 효율적인 학습

U-Net 전체를 미세조정하는 방식(815.45M)과 Transformer Block 부분(267.24M)만 미세조정하는 방식 그리고 self-attention layer(49.57M)만 미세조정하는 방식을 비교해보았을 때 큰 성능 차이를 보이지 않았다고 전함.

3) Mask-Free 학습 방법

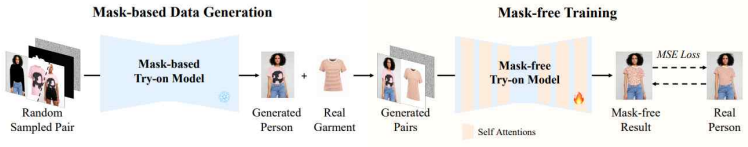


Figure 4: Overview of the mask-free training pipeline. We first use the trained mask-based model to generate synthetic person image from randomly sampled person-garment pairs. These synthetic person images, along with their corresponding original person and garment images, form the training data for the mask-free model.

단계	목적	절차
1단계	Mask-free Training Dataset 생성	입력: 원본 인물 + 의상 + 의류 mask 등 출력 : 학습에 사용할 Generated Person 생성
2단계	Mask-free 학습	입력 : 생성된 사람 이미지 + 실제 의류 이미지 출력 : 실제 사람 이미지

(7) 평가지표 정리

IS(Inception Score)

해당 metric의 의미만 보았을 때 생성된 이미지 데이터 셋 이 한 이미지 단위로는 뚜렷한 클래스를 가져야 하고 전체적으로는 골고루 다양한 클래스를 가져야 한다는 것을 KL Divergence로 정량화한 값.

$P(y=k|x)$: 개별 이미지가 클래스 k일 확률

$P(y=k)$ 전체 생성 이미지에서 클래스 k가 차이하는 비율

$$IS = \exp[Ex \sim P_G[D_{KL}(P(y=k|x) \parallel P(y=k))]]$$

Salimans, Tim, et al. "Improved techniques for training gans." Advances in neural information processing systems 29 (2016). https://proceedings.neurips.cc/paper_files/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html

FID

해당 metric의 의미만 보았을 때 두 개의 이미지 데이터 셋 분포 간 거리와 생김새 차이를 측정한 값을 의미함. InceptionNet 등을 활용하여 추출한 특징 벡터의 집합(임베딩 층)을 continuous multivariate gaussian으로 보고, 생성된 데이터 셋과 실제 데이터 셋 각각에 대해 평균과 공분산을 구하고 이후 Fréchet 거리를 계산함. 낮을수록 좋다.

μ_x: 실제 이미지 특징 평균, μ_g: 생성된 이미지 특징 평균, $Tr()$: 행렬 대각원소들의 합

Σ_x: 실제 이미지 특징 벡터들의 분산 Σ_g: 생성된 이미지 특징 벡터들의 분산

$$FID(x,g) = \|\mu_x - \mu_g\|_2^2 + Tr(\Sigma_x + \Sigma_g - 2\sqrt{(\Sigma_x \Sigma_g)})$$

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium https://arxiv.org/abs/1706.08500

KID

해당 metric의 의미만 보았을 때 두 개의 이미지 데이터 셋 분포 간 거리를 측정한 값을 의미함. MMD를 사용하고 커널 종류에 따라 민감도가 달라 질 수 있음. MMD 중 unbiased 추정식이 있어 샘플 수가 적어도 평균적으로 정확함. 정규분포를 가정하지 않았기 때문에 비교적 안정적. 낮을수록 좋다.

X: 실제 이미지 데이터 셋, Y: 생성된 이미지 데이터 셋, $X = \{x_1, \cdots, x_m\}$, $Y = \{y_1, \cdots, y_m\}$

$$KID = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)$$

Bińkowski, Mikołaj, et al. "Demystifying mmd gans." arXiv preprint arXiv:1801.01401 (2018). https://arxiv.org/abs/1801.01401

Gretton, Arthur, et al. "A kernel two-sample test." The Journal of Machine Learning Research 13.1 (2012): 723-773. https://jmlr.csail.mit.edu/papers/v13/gretton12a.html

SSIM (Structural Similarity Index Map)

인간의 눈이 밝기(Luminance), 대비(Contrast) 그리고 구조(Structural)에 민감하다는 사실을 가지고 각각 비교한 뒤 곱셈으로 종합하는 방식, 낮을수록 좋다.

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}$$

Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13.4 (2004): 600-612. https://ieeexplore.ieee.org/abstract/document/1284395

LPIPS(Learned Perceptual Image Patch Similarity)

사람의 지각(perception)과 더 잘 맞는다고 판단되는 함수를 학습을 통해 얻어 내는 방식을 사용함. 실제 이미지(x)와 생성된 이미지(x_0)를 CNN 기반의 네트워크를 사용하며 layer 단계별 feature map을 추출하여 원본과 생성된 이미지의 feature map($\hat{y}$, $\hat{y}_0$) 간 거리를 구하는 방식. 이를 perceptual similarity(인지적 유사도)라고 도 함.

$$d(x,x_0) = \sum_l \frac{1}{H_l, W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

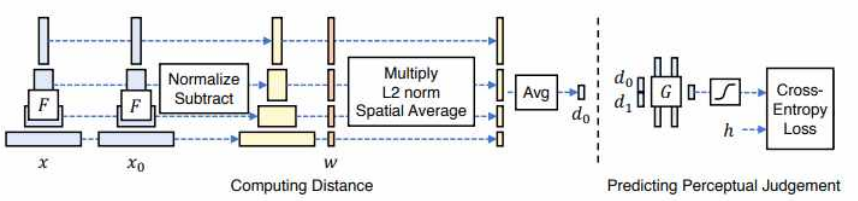


Figure 3: **Computing distance from a network** (Left) To compute a distance d_0 between two patches, x, x_0, given a network $\mathcal{F}$, we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector w, and take the ℓ_2 distance. We then average across spatial dimension and across all layers. (Right) A small network $\mathcal{G}$ is trained to predict perceptual judgment h from distance pair (d_0, d_1).

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric https://arxiv.org/abs/1801.03924

DreamSim(LPIPS의 확장)

ViT 백본을 사용하여 글로벌 코사인 유사도를 계산하는 방식. 각 백본의 class token을 이어 붙여 global cosine similarity를 사용해 한번에 유사도를 측정. 차별점은 DINO, CLIP 그리고 OpenCLIP 임베딩을 추가 정보로 활용하여 최종 거리를 계산함으로써 색상 텍스처 뿐만 아니라 의미적 차원까지 폭넓게 평가 가능하다고 주장.

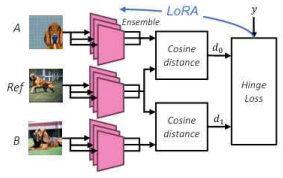
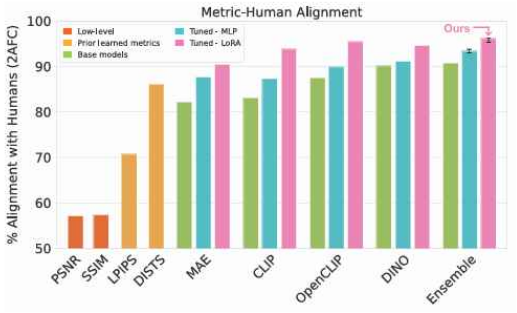


Figure 3: **Training overview.** Our perceptual metric is an ensemble of backbones, concatenating the representations together and fine-tuning with LoRA. We evaluate how well cosine distance on candidate feature spaces aligns with perceptual judgments y.

Objective. To better align a perceptual metric with human judgments, given triplet $(x, \tilde{x}_0, \tilde{x}_1)$ we maximize the difference between the perceptual distances d_0, d_1, with smaller distance associated with the distortion y humans voted to be most similar. To accomplish this, we map $y \in \{0, 1\} \rightarrow \hat{y} \in \{-1, 1\}$ and use a hinge loss, equivalent to triplet loss [14] between the embeddings, with a margin of $m = 0.05$:

$$\mathcal{L}(y, \hat{y}) = \max(0, m - \Delta d \cdot \hat{y}), \text{where } \Delta d = d_0 - d_1. \quad (2)$$



Fu, Stephanie, et al. "Dreamsim: Learning new dimensions of human visual similarity using synthetic data." arXiv preprint arXiv:2306.09344 (2023). https://arxiv.org/abs/2306.09344

(8) 사이즈 관련 평가 지표 정리

Size Does Matter :

생성된 이미지의 품질과 실제 착용감을 평가하기 위해 자원자를 받아 사람이 직접 옷을 입어보는 방식으로 평가함.

Table 2. The user study shows that COTTON is the most photo-realistic and remains the most human and clothing characteristics.

Method	PASTA-GAN	VITON-HD	HR-VITON	COTTON (Ours)
Photo-realistic	9.08%	7.56%	29.82%	**53.54%**
Try-on accuracy	8.66%	4.62%	27.40%	**59.32%**

Chen, Chieh-Yun, et al. "Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network." Proceedings of the IEEE/CVF international conference on computer vision. 2023
https://openaccess.thecvf.com/content/ICCV2023/papers/Chen_Size_Does_Matter_Size-aware_Virtual_Try-on_via_Clothing-oriented_Transformation_Try-on_ICCV_2023_paper.pdf

Size Variable VTON :

생성된 이미지상에 모델이 입고 있는 옷을 자연스럽게 분리한 후 주름의 정도에 따라 보정치를 곱해 최대한 편 상태에서 실측치(cm) 구하고 실제 옷과 비교하는 방식으로 평가함. 그 외에도 소매나 밑단의 길이만 관심 영역으로 두고 길이를 비교하는 방식도 있음.

1) Size Measurement

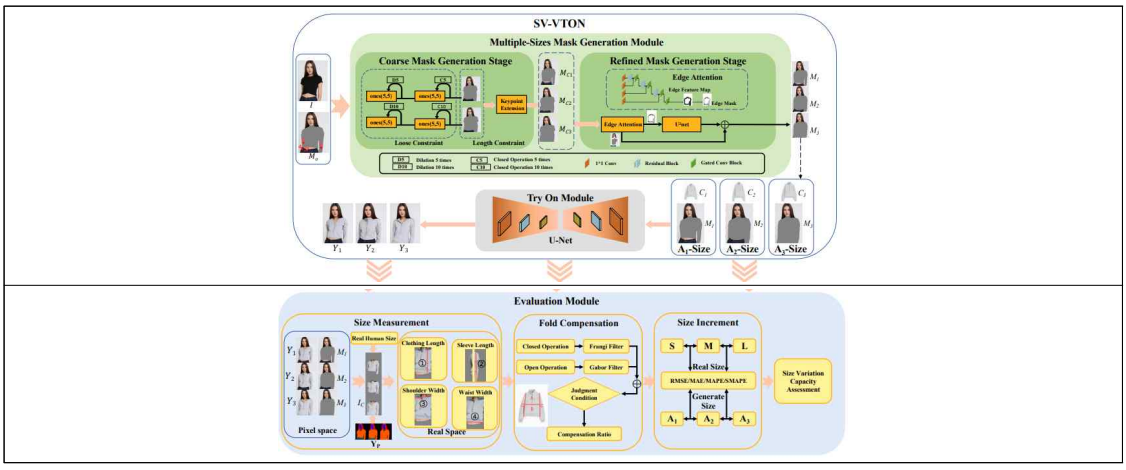
예측 결과와 정제 마스트를 통해 의류 영역(I_c)를 추출하여 네 가지 치수를 픽셀 단위로 계산하고 이후 픽셀 단위에서 cm로 변환하는 방식을 사용함.

2) Fold Compensation

주름에 정도에 따라 보정 계수 적용하여 길이 보정

3) Size Increment

실제 치수와 생성 치수를 MAE, RMSE, MAPE로 정량 평가. 오차가 작을수록 제대로 키웠다 평가.



Diffusion Model-Based Size Variable Virtual Try-On Technology and Evaluation Method https://arxiv.org/pdf/2504.00562v1
Size-Variable Virtual Try-On with Physical Clothes Size https://arxiv.org/pdf/2412.06201v1