Clothed Human Modeling

2025년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
정윤성

Outline

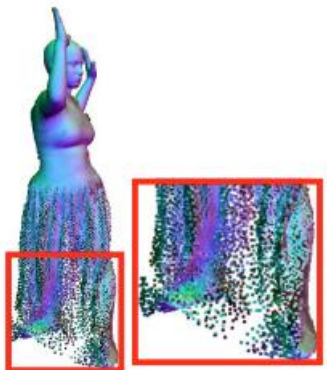
- Background
  - Human Shape
- ShapeBoost: Boosting Human Shape Estimation with Part-Based Parameterization and Clothing-Preserve[1)]
  - AAAI 2024
- FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling[2)]
  - CVPR 2025

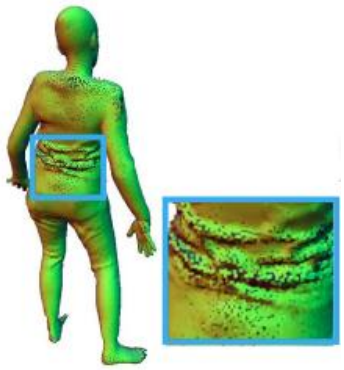
Background

- 3D 인체 복원 기술
  - Model-based
    - SMPL과 같이 사람 몸 모델의 파라미터를 맞춰서 형태 예측
      - 최적화 방식
        - ✓예측된 3D를 2D로 projection한 값과 2D GT를 비교해서 오차를 줄임
      - 회귀 방식
        - ✓이미지를 입력으로 주고 인체 파라미터를 예측
  - Model-free
    - 사람 모델 없이 이미지에서 바로 3D mesh의 각 vertex 위치를 직접 예측
    - Pose는 잘 예측하지만 shape는 예측이 떨어짐
      - 기본적인 인간 체형의 템플릿이 없어 수천 개의 3D vertex위치를 예측하기 어려움
      - 이때 점들이 연결되어 매끄러운 인간의 몸을 이룬다는 보장이 어려움

Background

- 3D 인체 복원 기술
  - 기존 연구는 SMPL과 같은 parametric Model에 크게 의존
  - SMPL 파라미터로 체형을 생성하면 몸에 딱 붙는 옷의 경우 잘 표현하지만 치마나 겉옷의 경우 표현이 제한적
    - 점의 개수나 연결 구조를 바꿀 수 없기 때문에 몸에서 떨어진 공간을 새로 만들 수 없음
      - 옷이 몸의 일부처럼 덩어리지는 현상
      - 치마 부분이 찢어지는 현상
      - 주름의 표현이 불가능한 현상





1) BIAN, Siyuan, et al. ShapeBoost: Boosting human shape estimation with part-based parameterization and clothing-preserving augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence.

ShapeBoost: Boosting Human Shape Estimation with Part-Based Parameterization and Clothing-Preserve[1)]

AAAI 2024

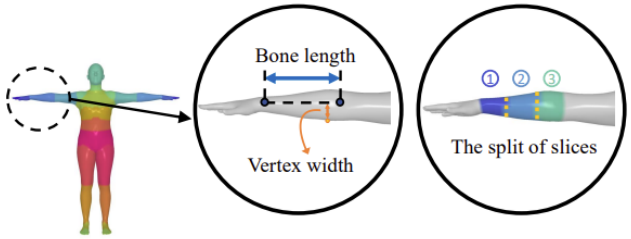
Introduction

- 단일 이미지로 3D 인체를 복원하는 것은 사람마다 체형과 의상이 다양하기 때문에 어려움
  - 기존 방식은 몸 전체를 PCA기반의 파라미터 (β)로 표현
  - 이 방식의 경우 다양한 체형(특히 극단적 체형)의 dataset 부족하기 때문에 부정확한 추론

- 기존의 해결 방식
  - 가상의 3D 모델로 합성 데이터를 만들어 훈련하는 방식
    - 한계: 합성 데이터는 옷의 질감이나 주름이 비현실적
      두꺼운 옷을 입거나 가려짐이 발생하는 경우 성능이 저하
  - Weak supervised learning
    - Shape cue 사용 (2D 실루엣, edge, 신체 분할 맵)
    - 한계: 옷의 종류나 자세에 따라 shape cue가 크게 변형
      여전히 극단적인 체형에서 부정확한 shape 추론

Method

- Part-based Parameterization
  - 기존의 shape를 β로 표현하는 방식은 몸 전체를 하나의 덩어리로 표현
    - 지역적인 디테일 부족, 각 부분별 제어에 한계
    - Shape Boost는 shape를 β 대신 뼈의 길이와 각 부위 단면 평균 너비로 표현
      - 뼈 길이는 keypoint로 계산, 너비는 신경망을 통해서 regression
  - 우선 SMPL mesh를 LBS weight에 따라 J=24개의 신체 부위로 구분
  - 각 부위의 중심 뼈는 양 끝 관절을 이은 직선으로 정의
    - (위팔의 경우 어깨 관절과 팔꿈치 관절을 직선으로 연결한 선을 중심 뼈로 가짐)
    - 피부의 한 점에서 뼈까지의 거리를 width로 가정
  - 각 신체부위는 뼈 방향으로 n개의 component로 분할
    - 각 slices에 있는 vertex의 평균 width가 해당 부위 두께



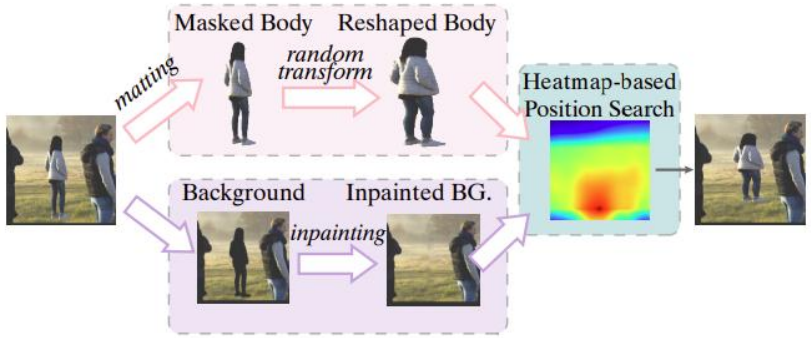
<l,w>

Method

- Semi-analytical Algorithm
  - 뼈의 길이 l과 너비 w가 주어졌을 때, 기본 자세를 취한 모델 T를 만드는 함수
  - T = M(l, w)
    - l: 각 뼈의 길이, w: 각 신체 부위 slices의 두께, T: T-pose mesh
    - 신경망을 사용해서 M을 도출하면 overfitting 발생
  - T = M(l, w) = MLP(M_0(l,w),l,w,Δl,Δw)
    - M_0; , rule based 방식으로 변형한 mesh, 현재 l, w, 차이 값을 입력받아 MLP로 예측
  - V = P(θ,M(l, w))
    - SMPL 모델의 자세 변형 함수 P를 통해서 원하는 pose로 mesh 변환

Method

- Clothing-Preserving Data Augmentation
  - 이미지 전체를 변형시키면 배경도 같이 변형되어 인체 mesh 데이터 증강의 의미가 사라짐
  - segmentation으로 이미지에서 사람을 분리
  - 배경을 inpainting
  - 분리한 사람 이미지를 affine transforming을 통해서 무작위 변형
    - 이때 camera 및 pose parameter가 주어지면 뼈를 일관되게 늘려 2D joint projection을 보장하기 때문에 변환 후의 뼈 길이를 얻을 수 있음
  - 변경된 human body를 inpainting된 배경 이미지에 다시 붙여 넣음
    - 이런 방식으로 옷의 질감, 주름, 조명, 배경 등의 조건이 현실적이고 다른 체형의 데이터셋을 생성



<데이터 증강>

Method

- 변형한 인체 치수에 맞게 GT 조절

  - $\overline{w_j} = \frac{\bar{s}}{s} \times \frac{ab \cdot l_j^{2D}}{\overline{l_j^{2D}}} \times w_j$

    - $\overline{w_j}$: 이미지 변형 후의 3D 너비, s: 카메라의 zoom 값, $\overline{l_j^{2D}}$: 이미지 변형 후의 2D 뼈 길이
    - 이때 문제는 $\bar{s}$의 정확한 값을 알기 어려움
      - 실제 3D 값을 supervised 대신 2D projection으로 값을 비교

  - $\overline{w}_k^{2D} = \frac{ab \cdot l_j^{2D}}{\bar{l}_j^{2D}} w_k^{2D}$

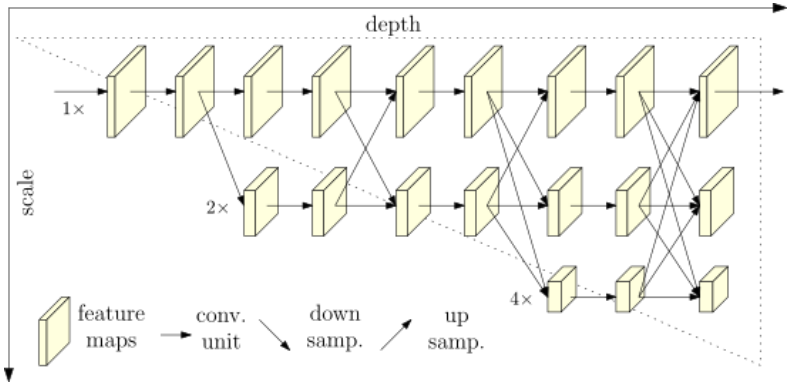
    - $\overline{w}$: 변형 후의 2D 너비, ab: scaling factor

Method

- Loss
  - 최종 loss: $L = L_{\text{pose}} + \mu_2 L_{\text{decomp}} + \mu_3 L_{\text{shape}}$
    - $\mu 1 = 0.01, \mu 2 = 0.1, \mu 3 = 1$
  - Shape Loss
    - $L_{\text{shape}} = \sum_j^J |\widehat{w_j}^{2D} - \overline{w_j}^{2D}|_2^2 + \mu_0 \sum_k^K |\widehat{w_j}^{2D} - \overline{w_j}^{2D}|_2^2$
    - w_j : 신체 부위 평균 너비, w_k : 신체 부위 특정 vertex의 너비
    - 부위 전체의 평균적인 두께와 특정 한 점이 중심 뼈에서 떨어진 거리를 모두 고려
    - HRNet 사용 이유: 이미지의 고해상도 정보를 끝까지 유지하면서 특징을 추출
  - Pose Loss
    - HybrIK의 Loss를 그대로 사용
    - 3D 관절 위치, 2D 관절 위치, SMPL 파라미터를 GT와 비교

Method

- Loss
  - Shape-decompose Loss
    - HRNet의 예측 값을 pseudo label로 사용
    - $L_{\text{decomp}} = L_{\text{bone}} + L_{\text{width}} + \mu_1 L_{\text{reg}}$
      - $L_{\text{bone}} = \sum_j^J (|\widetilde{x_j} - \widehat{x_j}|_1 + |\widetilde{l_j} - \widehat{l_j}|_1)$
        - ✓ $\widetilde{x_j}$: MLP가 예측한 3D 관절 위치, $\widetilde{l_j}$: 3D 뼈 길이
        - ✓ $\widehat{x_j}$: HRNet이 예측 한 3D 관절 위치, $\widehat{l_j}$: 3D 뼈 길이
      - $L_{\text{width}} = \sum_j^J \left(|\widetilde{w_j} - \widehat{w_j}|_2^2 + |\frac{\widetilde{w_j}}{\widetilde{l_j}} - \frac{\widehat{w_j}}{\widehat{l_j}}|_2^2 \right)$
        - ✓ $\widetilde{w_j}$: MLP가 예측한 3D 너비, $\widetilde{l_j}$: 3D 뼈 길이
        - ✓ $\widehat{w_j}$: HRNet이 예측 한 3D 너비, $\widehat{l_j}$: 3D 뼈 길이
      - $L_{\text{reg}} = |\tilde{\beta}|_2^2$
        - ✓ β의 값이 너무 커져 지나치게 비현실적인 체형이 나오는 것을 방지



<HRNet>

Experiments

- 정량 평가
  - PVE
    - 3D 모델 표면의 vertex 오차를 측정하는 지표
  - **SSP-3D** dataset 사용
    - 다양한 체형과 자세를 담은 운동선수의 이미지 311개로 구성된 평가 데이터
  - Table 2는 HBW(Humans in the Wild)라는 실제 야외 환경에서 찍은 데이터셋
  - 각각 키(Height), 가슴 둘레(Chest), 허리 둘레(Waist), 엉덩이 둘레(Hips)의 오차

Method	Model	PVE-T-SC ↓
HMR (Kanazawa et al. 2018)	SMPL	22.9
SPIN (Kolotouros et al. 2019)	SMPL	22.2
(Sengupta et al. 2020)	SMPL	15.9
(Sengupta et al. 2021b) †	SMPL	13.3
(Sengupta et al. 2021a)	SMPL	13.6
HybrIK (Li et al. 2021)	SMPL	22.8
LVD (Corona et al. 2022)	SMPL	26.1
CLIFF (Li et al. 2022b)	SMPL	18.4
SHAPY (Choutas et al. 2022)	SMPL-X	19.2
SoY (Sarkar et al. 2023)	SMPL	15.8
(Ma et al. 2023)	SMPL	18.8
(Sengupta et al. 2021a)*	SMPL	15.4
SHAPY (Choutas et al. 2022)*	SMPL	12.2
ShapeBoost (Ours)	SMPL	**11.4**
ShapeBoost (Ours)	SMPL-X	12.0

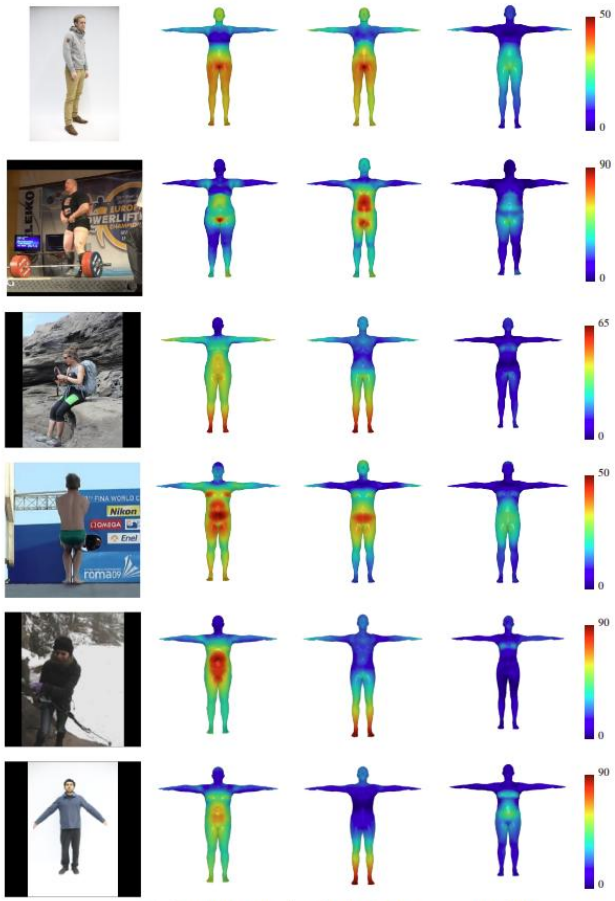
<Table 1>

Method	H	C	W	HC	$P2P_{20K}$
SPIN	59	92	78	101	29
Sengupta et al. 2020	135	167	145	102	47
TUCH	58	89	75	57	26
Sengupta et al. 2021a	82	133	107	63	32
CLIFF	-	-	-	-	27
SHAPY	**51**	65	69	57	**21**
ShapeBoost (SMPL)	66	**63**	58	**47**	25
ShapeBoost (SMPL-X)	68	69	**56**	49	22

<Table 2>

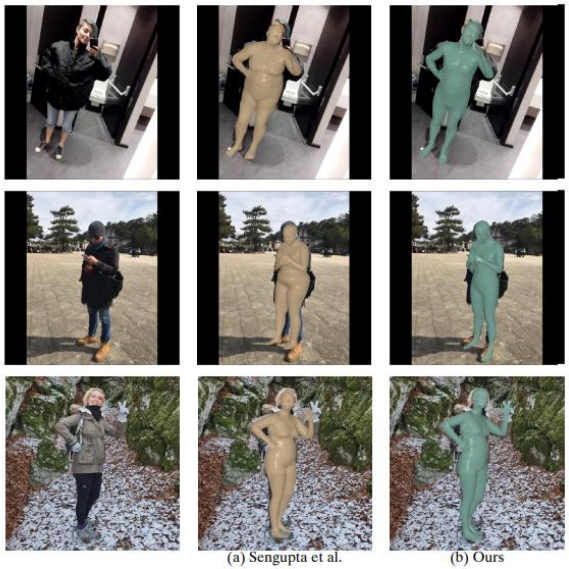
Experiments

- 정성 평가
  - 신체 어떤 부위를 잘 예측하는지 – 파란색이 정답에 가까움



(a) Sengupta et al. (b) SHAPY (c) Ours

(a) Sengupta et al. (b) Ours



(a) Sengupta et al. (b) Ours

Limitation

- 결국 최종 결과물이 SMPL의 β 파라미터로 표현
  - SMPL이 표현할 수 없는 체형은 mesh 복원이 어려움
- 옷 안의 shape 추정에만 초점
  - 옷 자체를 3D 모델로 생성하는 것은 아님
- 의류 데이터 증강시 단순 2D affine transform에 의존
  - 실제 비선형적인 체형 변화를 반영하지는 못함

1) YE, Hang, et al. FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. p. 15987-15997.

FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling[1)]

CVPR, 2025.

Introduction

- Animation이 적용된 Human avatar를 구현하기 위해서는 자세에 따라 자연스럽게 변형되는 의상을 정확히 모델링하는 것이 필수적
  - 하지만 한 장의 사진으로부터 옷을 입고 있는 사람의 3D 모델을 만들기가 어려움

- 기존 연구들의 경우 SMPL과 같은 파라미터 모델에 의존
  - Canonical Space에서 옷의 변형을 예측한 뒤 Linear Blend Skinning(LBS)를 사용해 목표 자세로 변형
  - 이 방식의 문제는 SMPL 모델이 고정된 Topology
    - Pose or shape에 의해 표면이 늘어나거나 줄어들 수는 있어도 vertex 개수는 고정
    - 따라서 헐렁한 옷의 경우 하나의 덩어리로 몸에 붙어 표현되거나 찢어짐 문제 발생

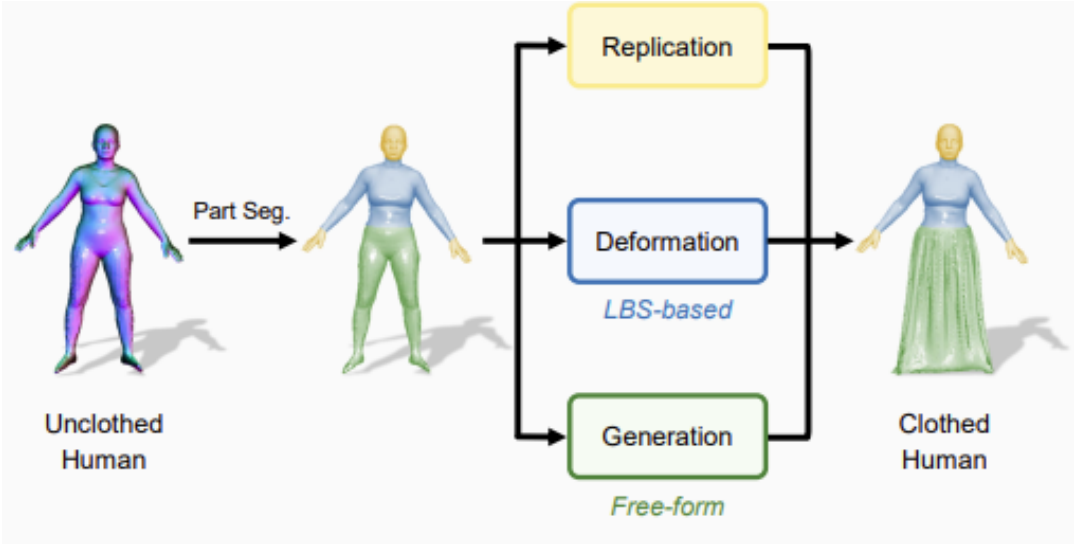
Introduction

- LBS는 몸에 가까운 부분에서는 적합한 변형 방식이지만 모든 영역에 LBS를 의존하는 것은 오히려 표현력을 떨어뜨림

- 이를 해결하기 위해 인체 표면을 세 가지 유형으로 분할하여 각기 다른 방식(Clothing-Cut Map)을 사용
  - 노출된 부분 (Unclothed, Yellow)
    - 변형이 따로 필요 없기 때문에 복제하여 사용
  - 몸에 붙는 부분 (Deformed, Blue)
    - 기존 LBS 기반 변형을 사용
  - 헐렁한 부분 (Generated, Green)
    - LBS로는 표현이 불가능하기 때문에 Free-form Generator 사용
      즉 Diffusion Model을 사용

Introduction

- Pipeline
  - Coarse Body Reconstruction
    - ICON을 사용하여 2D 입력 이미지(I)로 SMPL parameter(θ, β), 3D 몸통 mesh(B) 추출
  - Hybrid Clothing Generation
    - SMPL 파라미터를 3가지 영역으로 분할하고 각 영역에 맞는 방식을 적용

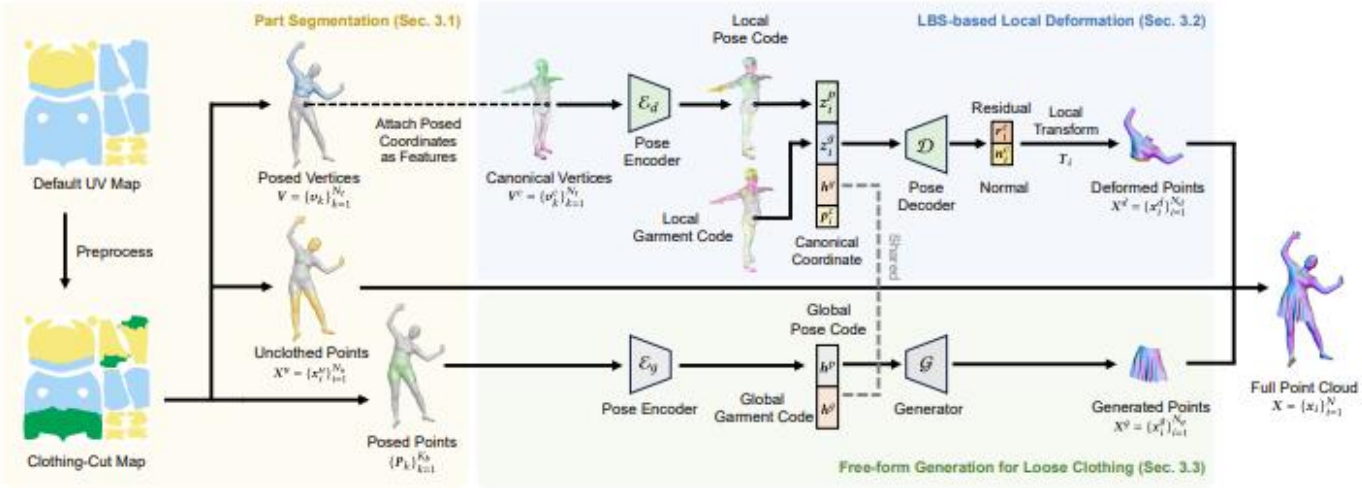


Method

- Human Part Segmentation
  - 세 영역(Unclothed, Deformed, Generated)으로 나눈 Clothing-Cut Map를 생성
    - Clothing-Cut Map 생성 시 Segment Anything Model(SAM)을 사용
  - Clothing-Cut Map 생성 과정
    - 헐렁한 부분(Generated, Green)
      - 전후방 2D Normal Map Rendering
        - ✓2D Normal map은 옷의 질감, 조명 등의 조건과 상관없이 주름과 굴곡과 같은 기하학적 정보만 존재
        - ✓SAM에 입력해서 2D에서 헐렁한 부분 분리
      - 2D segmentation
        - ✓주름이 많거나 헐렁한 부분을 찾아 2D mask 생성
      - 3D Mapping
        - ✓2D mask 영역을 3D 공간으로 역투영
        - ✓이를 통해 헐렁한 부분을 찾아냄

Method

- Human Part Segmentation
  - Clothing-Cut Map 생성 과정
    - 노출된 부분(Unclothed, Yellow)
      - 피부 영역 사전에 정의(머리, 손, 발)
      - 헐렁한 옷이 덮지 않는 나머지 노출 부위도 노출된 부분으로 정의
      - 노출된 부분은 추가 변형 없이 복제하여 사용
    - 몸에 붙는 부분 (Deformed, Blue)
      - 앞서 확정된 영역을 제외하고 남는 나머지 모든 부분

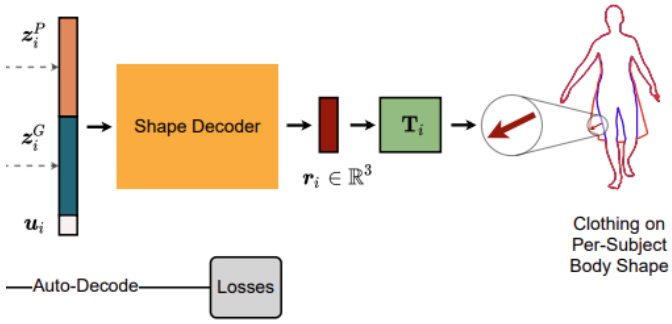


Method

- LBS-based Local Deformation
  - 몸에 붙는 부분 (Deformed, Blue)을 처리할 때 사용
  - 코드 추출
    - Local Pose Code
      - $\{\phi_k^p\}_{k=1}^{N_t} = E_d(V^c, V)$.
        - ✓V: 현재 자세의 3D mesh, V^c: T-pose의 3D mesh
      - PointNet++ 을 사용하여 각 vertex의 자세 변화 정보 추출
      - ϕ_k^p : 각 부위별 local pose code
    - Garment Code
      - 자세와는 상관없이 표준 자세(T-pose)를 기준으로 학습
      - Local Garment Code
        - ✓각 vertex가 옷의 어떤 부분에 속하는지를 확인
      - Global Garment Code
        - ✓입고 있는 옷의 종류가 무엇인지를 확인
        - ✓옷 종류마다 고유한 코드값을 부여하고 그 값을 바탕으로 옷을 복원

Method

- LBS-based Local Deformation
  - Pose Decoder로 POP에 사용된 MLP기반의 decoder 구조를 그대로 사용
  - $[r_i^c, n_i^c] = D(z_i^p, z_i^g, h_g, p_i^c)$
    - 입력: [Local Pose Code, Local Garment Code, Global Garment Code, Canonical Coordinate]
      - T-Pose에서 현재 자세로의 변화량, 옷의 부위 정보, 옷 종류 정보, T-pose에서의 3D vertex 위치
    - 출력: T-Pose에서의 변형량(옷이 피부에서 얼마나 떨어져야 하는지)과 normal vector
  - 표준 자세에서 현재 자세로 변형하기 전에 자세 변화에 따른 신체, 옷의 변형을 미리 반영
- LBS 변형
  - 앞서 Point Net으로 구한 자세 변화를 반영하여 현재 자세 모양으로 인체를 변형
  - POP 의 decoder는 8층으로 된 MLP로 활성화 함수는 ReLU사용



<POP Decoder>

Method

- Free-form Generation for Loose Clothing
  - 헐렁한 부분 (Generated, Green)을 처리할 때 사용
  - Structure-aware Pose Encoding
    - 3D mesh를 K_b개의 신체 부위로 나누고 각 부위에서 점을 sampling
      - (몸통,위팔, 아래팔, 손, 허벅지, 종아리, 발, 머리)
    - SpareNet의 인코더를 PointNet에서 PointNet ++로 변경해 사용
      - PointNet++를 사용해서 sampling된 신체 부위별로 local features를 추출
        - ✓PointNet의 경우 몸 전체 point cloud를 한번에 보고 global한 자세 정보를 추출
        - ✓PointNet++의 경우 각 부위별로 local 자세 정보를 추출
    - Local features들 max-pooling을 사용해 부위 기반 자세 코드를 생성
    - 가장 두드러지는 움직임 변화만 선택
  - Free-form Generation
    - 부위 기반 자세 코드와 전역 의상 코드를 입력
    - MLP를 통해 헐렁한 옷 부분을 point cloud로 생성

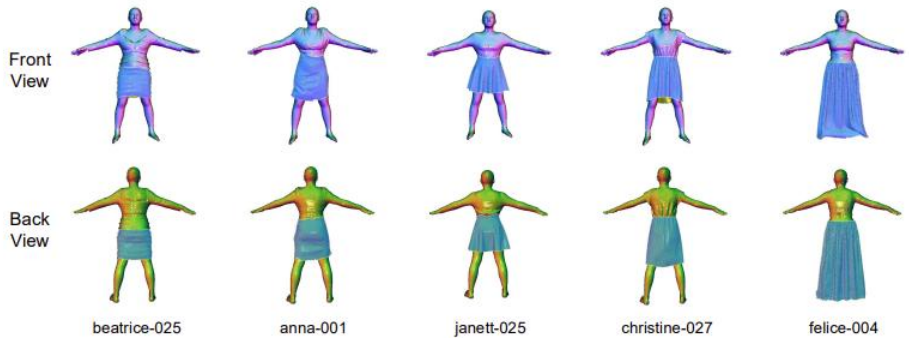
Method

- Training
  - 전체 loss
    - $L = \lambda_{cd}L_{cd} + \lambda_n L_n + \lambda_{rd}L_{rd} + \lambda_{rg}L_{rg} + \lambda_{col}L_{col}$
  - $L_{cd} = \frac{1}{N}\sum_{i=1}^{N} \min_j |x_i - \hat{x_j}|_2^2 + \frac{1}{M}\sum_{j=1}^{M} \min_i |x_i - \hat{x_j}|_2^2$
    - 예측 point cloud (x_i)와 가장 가까운 gt point cloud($\hat{x_j}$)의 거리 제곱
    - 반대로 gt와 가장 가까운 예측 point cloud를 찾아 거리 제곱
  - $L_n = \frac{1}{N}\sum_{i=1}^{N} |n_i - \hat{n}\left(\arg\min_{\hat{x_j}} d\left(x_i, \hat{x_j}\right)\right)|_1$
    - 예측된 점에서 가장 가까운 gt를 찾아 normal vector끼리 비교
  - $L_{rd} = \frac{1}{N_d}\sum_{i=1}^{N_d} |r_i^c|_2^2$
    - 옷이 피부에서 너무 멀리 떨어지지 않도록
  - $L_{rg} = \sum_{i=1}^{N_d} \frac{1}{N_d} |z_i^g|_2^2 + |h^g|_2^2$
    - 의상 종류를 나타내는 코드가 너무 큰 수를 갖지 않도록 제한
  - $L_{col} = \frac{1}{N_g}\sum_{j=1}^{N_g} \max\{\epsilon - d\left(x_j^g\right), 0\}$
    - 옷이 몸 안으로 파고드는 것을 방지, ϵ의 경우 최소한의 마진을 의미

Experiments

- 정량 평가
  - 헐렁한 옷이 많이 포함된 ReSynth dataset으로 test 진행
    - Fréchet Inception Distance(FID)
      - Inception V3모델을 사용해서 사실적인지를 판단하는 지표
    - Mean Square Error(MSE)
      - Point cloud간의 차이가 아니라 normal map간의 비교

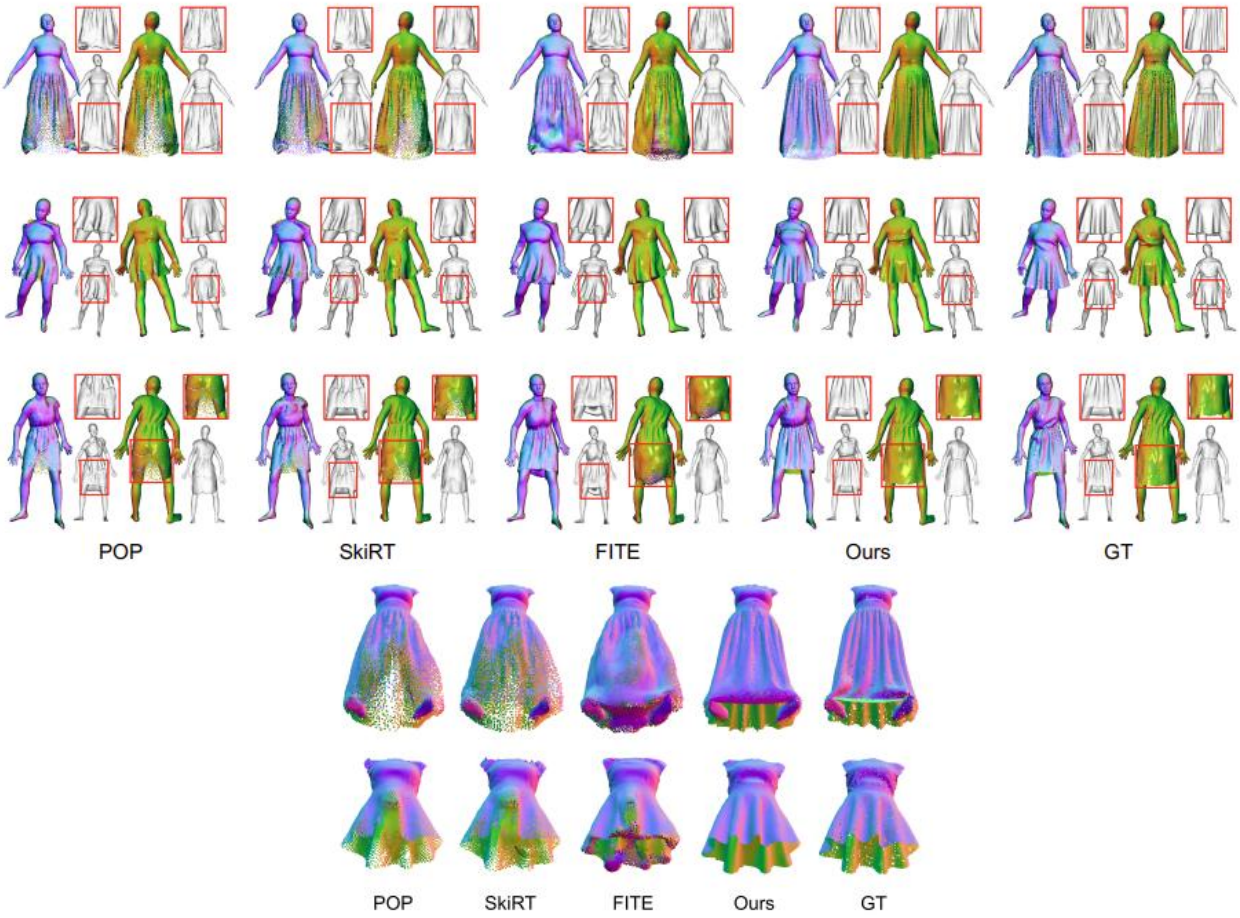
Subject	All		felice-004		janett-025		christine-027		anna-001		beatrice-025	
Metric	FID ↓	MSE ↓	FID ↓	MSE ↓	FID ↓	MSE ↓	FID ↓	MSE ↓	FID ↓	MSE ↓	FID ↓	MSE ↓
POP [39]	57.87	2.88	66.43	5.80	52.55	2.02	61.09	2.64	51.48	2.05	57.82	1.86
SkiRT [40]	53.32	2.72	63.27	5.70	48.23	2.03	55.84	2.44	50.26	**1.81**	54.00	**1.60**
FITE [33]	39.02	2.70	**38.61**	**5.09**	35.81	2.09	40.83	2.52	**38.21**	1.97	41.62	1.82
Ours	**37.75**	**2.61**	42.41	5.24	**27.95**	**1.92**	**37.43**	**2.35**	39.63	1.89	**41.24**	1.68



<Normal map dataset>

Experiments

- 정성 평가
  - 위에서부터 felice-004, janett-025, christine-027



감사합니다.