2026 동계 세미나

Human Mesh Recovery

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

Presented By
오정선

Outline

- Background
  - SMPL
  - Human Mesh Recovery
  - Projection

- TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation
  - CVPR 2024

- BLADE:Single-view Body Mesh Estimation through Accurate Depth Estimation
  - CVPR 2025

Background

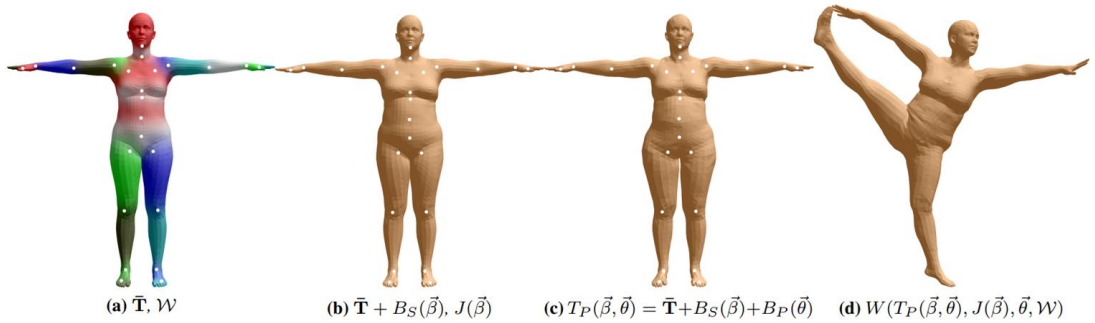
- SMPL
  - 개념
    - 인체를 체형(β) + 자세(θ) 파라미터로 표현하는 Parametric body model
    - Linear Blend Skinning (LBS) 사용
      - 관절 회전에 따라 메쉬를 부드럽게 변형시키는 방법
      - 물리적으로 일관된 인체 움직임 생성
    - HMR(Human Mesh Recovery) 분야에서 단일 이미지로부터 3D 인체를 복원하기 위한 표준 인체 표현
  - SMPL vs SMPL-X
    - SMPL
      - 전신 중심 인체 모델
      - 몸통·팔다리 포즈 표현에 초점
      - 기본 포즈만 지원
    - SMPL-X
      - SMPL을 확장한 통합 인체 모델
      - 몸 + 손 + 얼굴을 모델로 표현
      - 손가락 관절과 표정 포함

Background

- SMPL
  - Method
    - 입력
      - β : 체형 파라미터 ($\beta \in \mathbb{R}^{10}$)
      - θ : 자세 파라미터 ($\theta \in \mathbb{R}^{24x3}$)
    - 출력
      - N = 6890개 vertex로 구성된 인체 메쉬

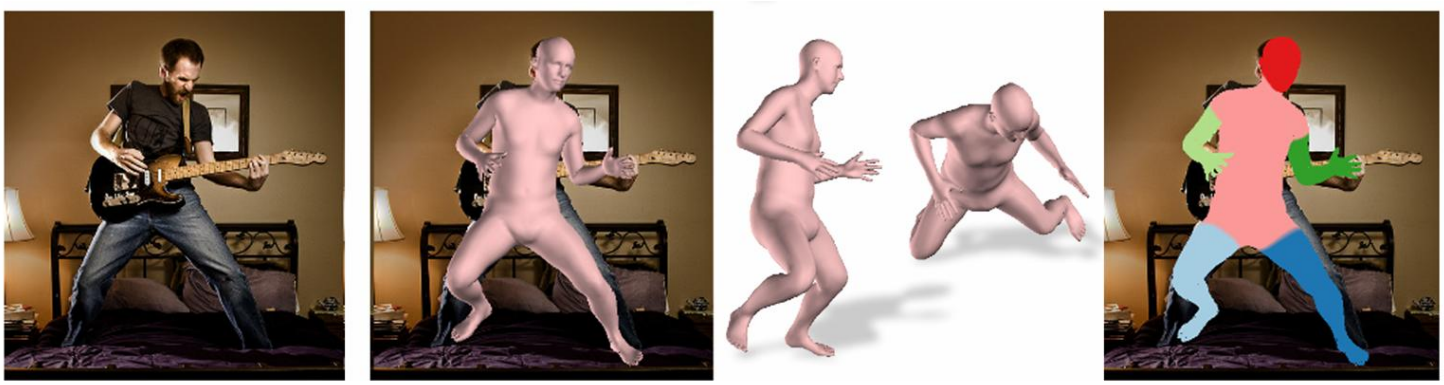
$$M(\beta, \theta) \in \mathbb{R}^{3xN}, \quad N = 6890$$

    - SMPL 흐름도
      - (a) 평균 인체 메쉬
      - (b) β 체형 파라미터 적용 (rest model 생성)
      - (c) θ 파라미터에 따른 체형 보정
      - (d) 최종 θ 파라미터에 따른 포즈 적용



Background

- Human Mesh Recovery
  - 개념
    - 단일 이미지에서 SMPL 파라미터를 직접 추정하는 end-to-end 프레임워크
    - 입력 이미지에서 2D 시각 정보 → 3D 메쉬 복원을 위해 직접 회귀 문제로 정의
  - 2D/3D 모호성 해결을 위해 Adversarial Prior 도입
    - SMPL 파라미터가 실제 인체 분포에 속하도록 제약
    - Pose / Shape 공간을 데이터 기반으로 regularization으로 안정화
  - Iterative Error Feedback(IEF) 구조 사용
    - Pose / Shape 공간을 데이터 기반으로 regularization으로 안정화
    - $\Theta_{t+1}=\Theta_t+\Delta\Theta$ (Iteration 설정 횟수만큼 수행)
    - 이미지 특징 + 현재 파라미터 추정값을 사용하여 메쉬 수렴



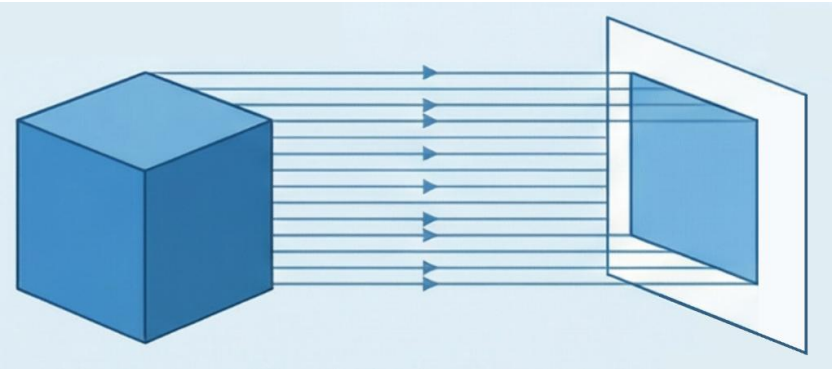
Background

- Projection
  - Orthogonal Projection
    - 3D 점 (x,y,z)에서 깊이 정보 z를 무시하고 2D 평면에 투영 $x = X, \quad y = Y$
  - Perspective Projection
    - 3D 점 (x,y,z)에서 카메라 중심 기준으로 2D 평면에 투영
    - 깊이 z에 따라 2D 위치가 비선형적으로 변함

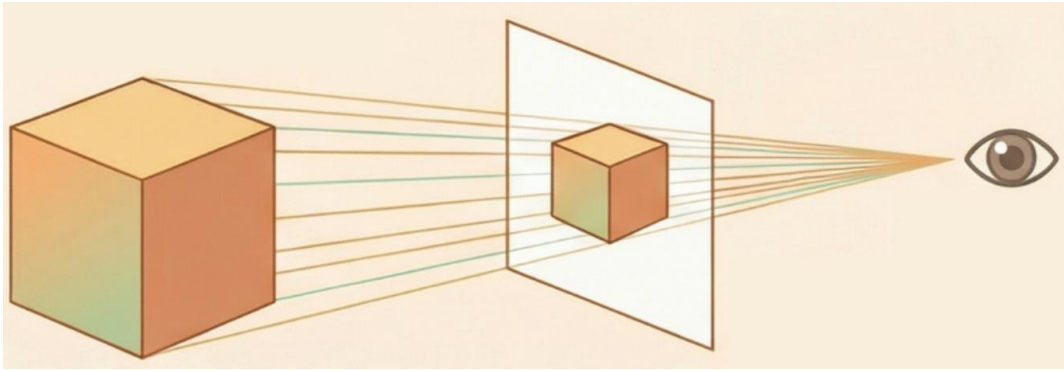
$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}$$

  - Weak-Orthogonal Projection
    - 전체 물체의 평균 깊이만 반영하여 선형으로 근사
    - 대부분 초기 HMR이 이 투영 방식을 사용

$$x = sX + t_x, \quad y = sY + t_y$$



<Orthogonal Projection>

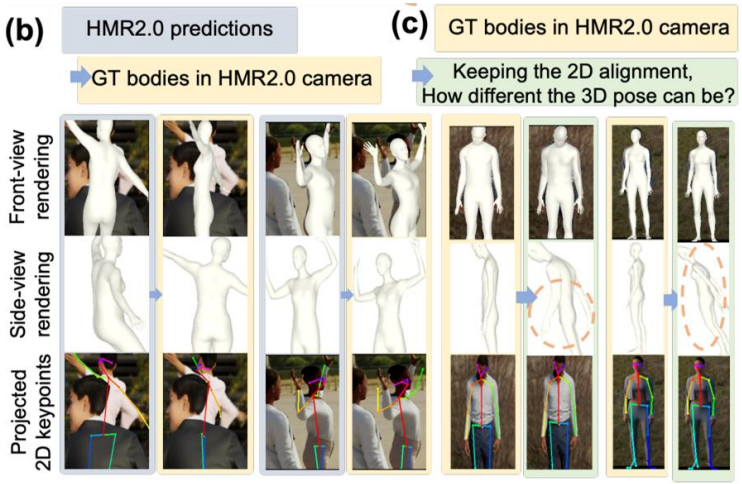


<Perspective Projection>

TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation [CVPR 2024]

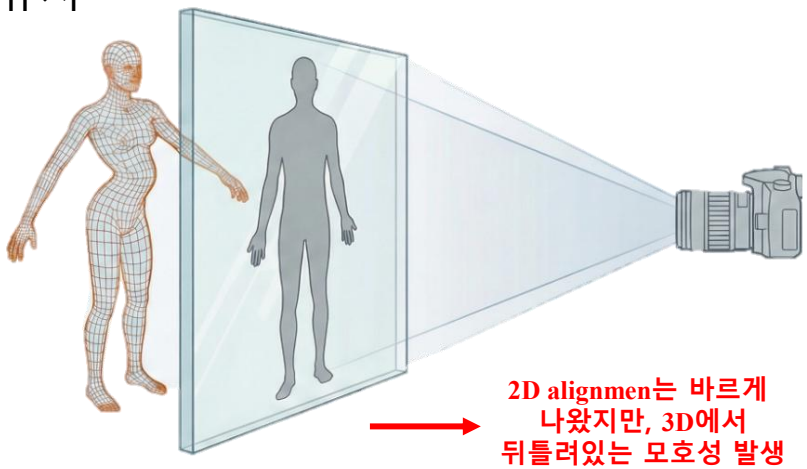
Introduction

- Motivation
  - 기존 HMR의 접근법
    - 3D ground-truth가 부족한 in-the-wild 환경에서 2D keypoint supervision과 3D pseudo-GT에 의존하여 학습
  - 문제 현상
    - 2D alignment 성능 ↑ 3D pose 정확도 ↓
      - Camera model mismatch로 인한 구조적 한계
  - 기존 HMR의 한계점
    - 기존 연구는 주로 데이터의 양과 품질에 집중
    - Supervision 자체가 3D를 왜곡할 수 있다는 점이 명확하게 분석되지 않음



Introduction

- Observation
  - 원인
    - Camera Model Bias
      - 대부분 Weak-perspective 또는 고정된 camera를 가정으로 접근
      - 그 결과 3D → 2D projection error 발생
    - Pseudo-GT Supervision
      - 2D Keypoints + 잘못된 camera 가정을 기반으로 생성
      - 잘못된 3D pose를 정답으로 학습
  - 목표
    - Camera parameter를 알 수 없는 상황에서 2D supervision의 장점을 유지하면서 3D pose 정확도를 유지



Method

- Threshold-Adaptive Loss Scaling: TALS
  - Pseudo-GT기반 3D pose loss를 조절하여 과도한 3D 최적화를 방지
  - 제안 방법
    - 학습 신호로서 더 이상 이득이 없는 오차 구간을 구분하기 위한 임계치 설정
      - 2D loss가 임계값보다 큰 경우 → 기존 방식과 동일하게 3D loss 적용
      - 2D loss가 임계값 이하인 경우 → 3D loss weight 감소
  - 3D Pose Loss (p-GT)
    - ε_θ : HMR2.0 예측 pose vs BEDLAM 3D GT pose

$$\mathcal{L}_{\theta p_{GT}} = \begin{cases} \|\theta - \theta_g\|^2 & \text{if } \mathcal{L}_{\theta p_{GT}} > \varepsilon_\theta \\ \alpha_\theta \cdot \|\theta - \theta_g\|^2 & otherwise \end{cases}$$

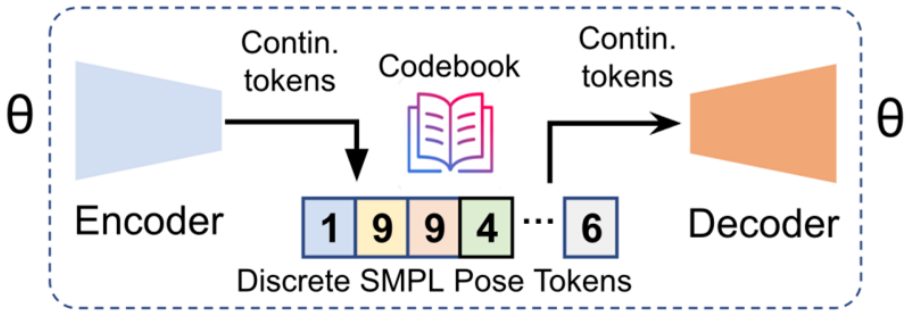
  - 2D Keypoints Loss
    - $\varepsilon_{J_{2D}}$: BEDLAM GT 3D pose → 2D Keypoints vs BEDLAM 2D GT keypoints

$$\mathcal{L}_{J_{2D}p_{GT}} = \begin{cases} \left|J_{2D} - J_{2D_g}\right| & \text{if } \mathcal{L}_{J_{2D}p_{GT}} > \varepsilon_{J_{2D}} \\ \alpha_{J_{2D}} \cdot \left|J_{2D} - J_{2D_g}\right| & otherwise \end{cases}$$

    - 효과
      - 작은 2D / p-GT 오차에 대한 과도한 최적화 방지
      - 학습 과정에서 3D pose 안정성 유지

Method

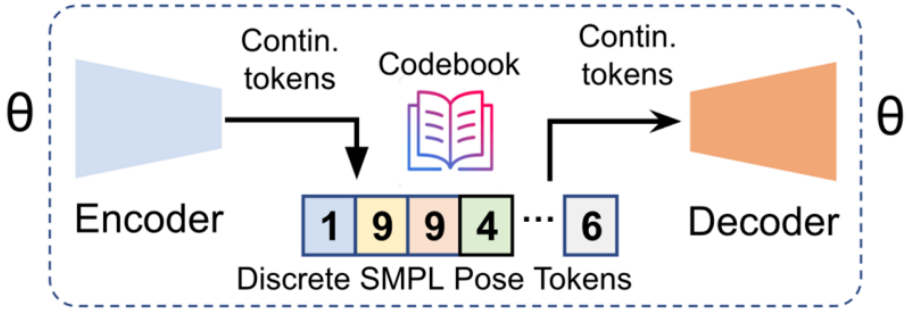
- Tokenization
  - Continuous Pose Regression의 한계
    - 기존 HMR는 관절을 연속 벡터로 회귀
    - 2D loss 강제 맞춤으로 인해 3D 왜곡 → 물리적으로 불가능한 pose발생
  - Tokenizer
    - Valid Pose Space를 학습하는 Prior 모델
  - Tokenizer 구조 (VQ-VAE)
    - SMPL body pose
      - $\theta = [\theta_1, \dots, \theta_{21}], \quad \theta_i \in \mathbb{R}^6$ 를 이산 토큰 시퀀스로 변환
    - Encoder
      - $z = E(\theta)$, pose를 연속 latent representation으로 압축



(a) Tokenization

Method

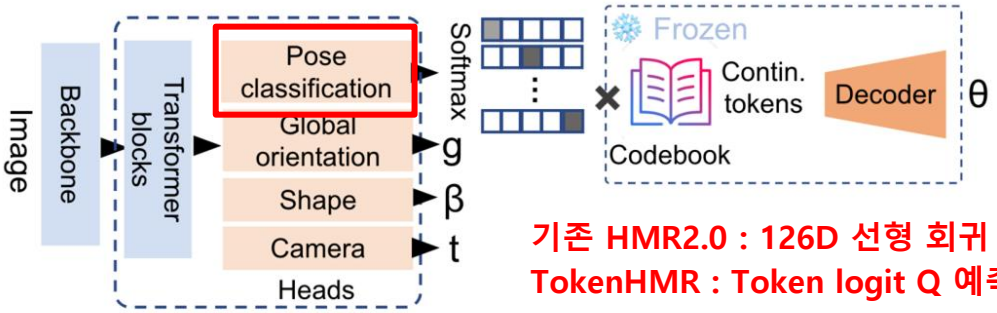
- Tokenization
  - Tokenizer 구조 (VQ-VAE)
    - Codebook (학습 가능한 자세 Vocabulary)
      - $CB = \{c_k\}_{k=1}^{K}$, AMASS 데이터 분포를 반영한 pose embedding 저장소
    - Quantization
      - $\hat{z_i} = \arg\min_{c_k \in CB} \|z_i - c_k\|_2$
      - Encoder가 출력한 연속 벡터 z_i를 Codebook 안에 가장 가까운 벡터로 치환
    - Decoder
      - Mocap에서 학습된 pose prior를 유지하기 위해 freeze
    - Overall Loss
      - $\mathcal{L}_{VQ} = \lambda_{\mathcal{RE}}\mathcal{L}_{\mathcal{RE}} + \lambda_{\mathcal{E}}\mathcal{L}_{\mathcal{E}} + \lambda_{\mathcal{C}}\mathcal{L}_{\mathcal{C}} = \lambda_{\mathcal{RE}}\mathcal{L}_{\mathcal{RE}} + \lambda_{\mathcal{E}}\|sg[z] - e\|_2 + \lambda_{\mathcal{C}}\|z - sg[e]\|_2$
      - Tokenizer가 원래 pose를 정확히 복원하도록 강제



(a) Tokenization

Method

- Architecture
  - Backbone
    - HMR2.0 Vison Transformer
  - Workflow
    - SMPL pose를 Body pose / Global orientation 으로 분리하여 추론
      - Body pose
        - ✓manifold 구조이기 때문에 Tokenization 적용
        - ✓Token logits Q 예측 → Softmax(Q) × Codebook → Tokenizer decoder → Pose
          - logit Q [B, M, K] : pose를 구성하기 위한 codebook 토큰 선택 점수
        - ✓Softmax 적용 이유
          - $\bar{z} = \sigma(\boldsymbol{Q}_{M\times K}) \times \boldsymbol{CB}_{K\times D} \approx \hat{z}$
          - 코드 선택을 연속 확률 분포로 바꾸어 학습을 가능하게 함

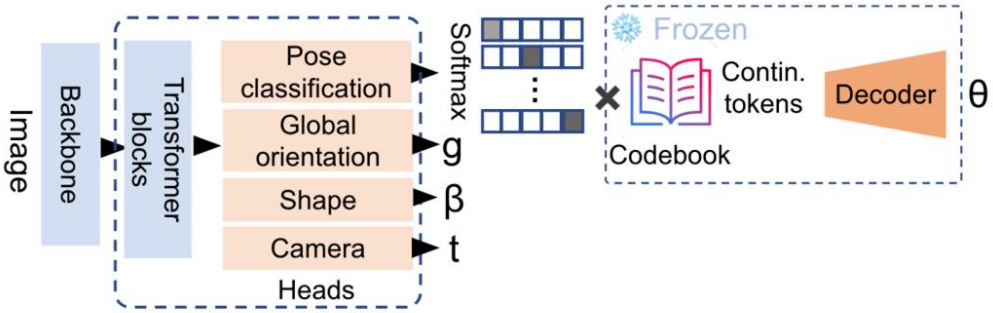


기존 HMR2.0 : 126D 선형 회귀
TokenHMR : Token logit Q 예측

(b) TokenHMR

Method

- Architecture
  - Workflow
    - SMPL pose를 Body pose / Global orientation 으로 분리하여 추론
      - Global orientation
        - ✓단순 회전이기 때문에 Linear Regression 적용
  - Loss
    - 3D GT Dataset Loss
      - $\mathcal{L}_{GT} = \lambda_\theta \mathcal{L}_\theta(\theta, \theta_g) + \lambda_\beta \mathcal{L}_\beta(\beta, \beta_g) + \lambda_{3D} \mathcal{L}_{3D}\left(J_{3D}, J_{3D_g}\right) + \lambda_{2D} \mathcal{L}_{2D}\left(J_{2D}, J_{2D_g}\right)$
    - 3D p-GT / 2D Dataset
      - TALS 적용
  - Total Loss
    - $\mathcal{L}_{Total} = \mathcal{L}_{GT} + \mathcal{L}_{\theta_{pGT}} + \mathcal{L}_{J_{2D}PGT}$



(b) TokenHMR

Experiments

- Quantitative Result

MVE : vertex 위치 평균 오차
MPJPE : 관절 위치 평균 오차
PA-MPJPE : 정렬 후 관절 오차

Training Datasets	Method	EMDB [23]			3DPW [55]		
		MVE	MPJPE	PA-MPJPE	MVE	MPJPE	PA-MPJPE
SD	HybrIK [31]	122.2	103.0	65.6	94.5	80.0	48.8
SD	CLIFF [32]	122.9	103.1	68.8	81.2	69.0	43.0
SD	HMR2.0 [12]	120.1	97.8	61.5	84.1	70.0	44.5
BL	BEDLAM-CLIFF [3]	113.2	97.1	61.3	85.0	72.0	46.6
BL	HMR2.0	106.6	90.7	51.3	88.4	72.2	45.1
BL	TokenHMR	104.2	88.1	49.8	86.0	70.5	43.8
SD + ITW	HMR2.0 [12]	140.6	118.5	79.3	94.4	81.3	54.3
SD + ITW	TokenHMR	124.4	102.4	67.5	88.1	76.2	49.3
SD + ITW + BL	HMR2.0	120.7	99.3	62.8	88.4	77.4	47.4
SD + ITW + BL	HMR2.0 + TALS	115.7	96.7	58.5	89.6	73.5	46.8
SD + ITW + BL	HMR2.0 + Token	116.1	95.6	62.2	86.6	75.0	48.0
SD + ITW + BL	HMR2.0 + TALS + VPoser [42]	116.8	97.9	56.4	87.1	73.7	45.7
SD + ITW + BL	TokenHMR	109.4	91.7	55.6	84.6	71.0	44.3

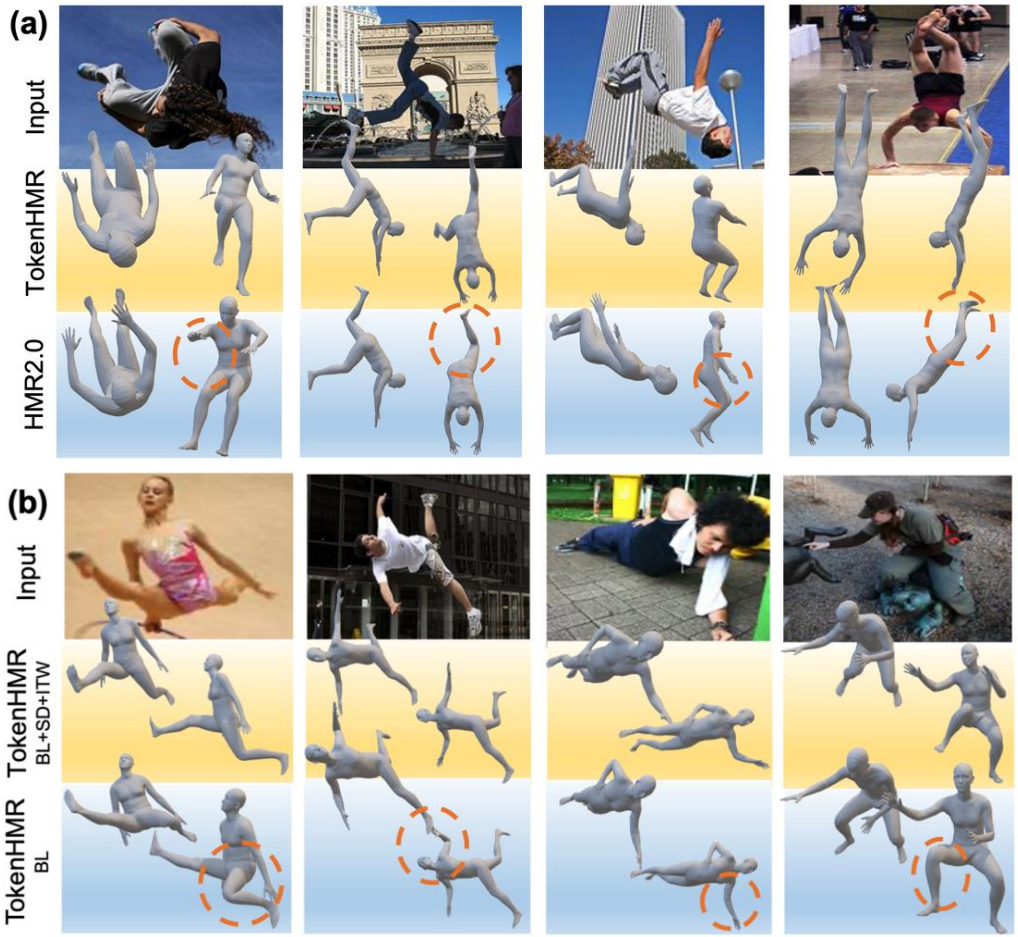
<SOTA 메소드와 학습 데이터 조합에 따른 성능 비교>

Method	Crop 30%			Crop 50%		
	MVE	MPJPE	PA-MPJPE	MVE	MPJPE	PA-MPJPE
HMR2.0 [12]	135.24 (+14.98)	113.39 (+14.13)	70.68 (+7.86)	166.71 (+46.45)	137.88 (+38.59)	90.30 (+27.48)
TokenHMR	124.09 **(+14.71)**	104.72 **(+13.01)**	62.13 **(+6.52)**	150.29 **(+40.91)**	125.99 **(+34.28)**	78.88 **(+23.27)**

<이미지 Cropping 환경에서 강건성 평가>

Experiments

- Qualitative Result



<SOTA 메소드와 정성적 비교>

Experiments

- Ablation Study
  - Tokenizer

학습 데이터

		AMASS [37]		MOYO [52]	
	Method	MVE ↓	MPJPE ↓	MVE ↓	MPJPE ↓
CB	1024 × 256	11.5	4.6	27.1	15.7
	2048 × 128	9.4	3.1	22.5	12.3
	2048 × 256	8.3	2.2	19.9	10.4
Tokens	80	12.5	4.1	24.4	16.7
	160	8.3	2.2	19.9	10.4
	320	8.1	1.9	19.0	10.1
Noise	Yes	8.3	2.2	19.9	10.4
	No	7.9	1.9	21.0	11.5
테스트 데이터	AMASS + MOYO*	8.7	2.6	16.5	7.6

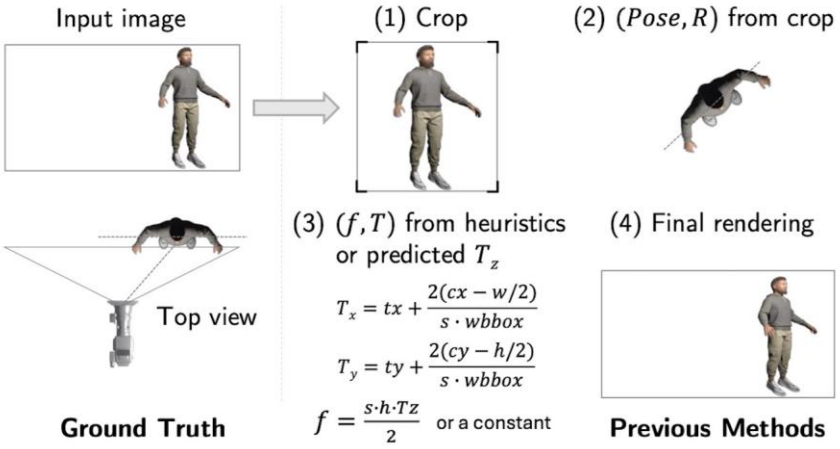
<Tokenizer Ablation study>

BLADE:Single-view Body Mesh Estimation through Accurate Depth Estimation

[CVPR 2025]

Introduction

- Motivation
  - 단일 이미지에서 인체 메쉬 + 카메라 파라미터 동시 추정 문제
  - 기존 HMR의 공통 가정
    - 사람이 카메라로부터 멀리 떨어져 있다는 가정 → Near-Orthograpic Projection 사용
    - Focal length 및 translation을 heuristic하게 설정
  - 실제 환경에서의 한계
    - Close-range 이미지에서 강한 Perspective distortion 발생
    - 3D pose 정확도와 2D alignment 간 trade-off 발생



Introduction

- Observation
  - Perspective distortion의 재정의
    - Perspective distortion은 focal length f 때문이 아니라 T_z에 의해 결정
  - 기존 방식의 오류
    - 왜곡의 원인을 focal length 추정 오류로 인식
    - Orthographic → perspective 변환에 heuristic을 사용하면서 오차 누적
  - 핵심 통찰
    - Heuristic 방식에 의존하지 않음
    - T_z를 분리하고 정확히 추정하면 pose, shape, camera 파라미터 문제를 해결 가능

Method

- 목표
  - Single image가 주어졌을 때 정확한 3D human mesh recovery & good 2D alignment
  - SMPL-X

$$SMPL-X\ vertices[x_m, y_m, z_m]\ as:\ [x, y, z] = [x_m, y_m, z_m] + [T_x, T_y, Tz]$$

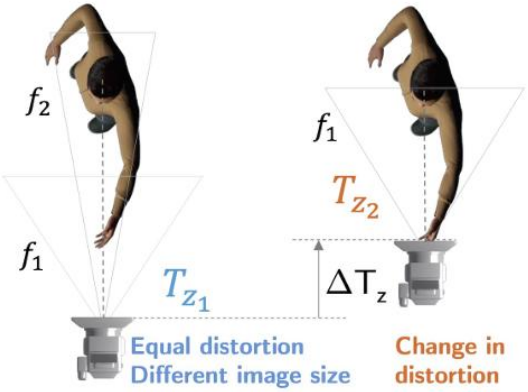
  - Perspective projection

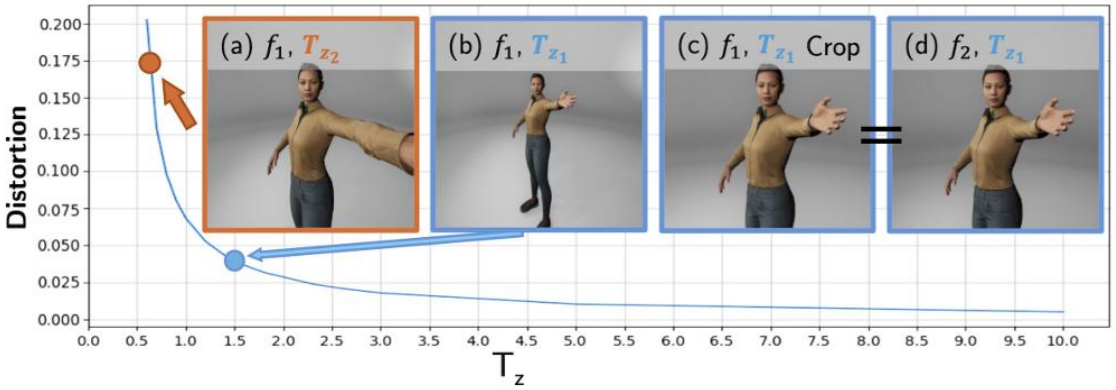
$$(u, v) = f\left(\frac{x}{z}, \frac{y}{z}\right) = f \cdot \left[\frac{x_m + T_x}{z_m + T_z}\right]$$

Linear Impact

Nonlinear Impact

  - T_z가 미치는 영향

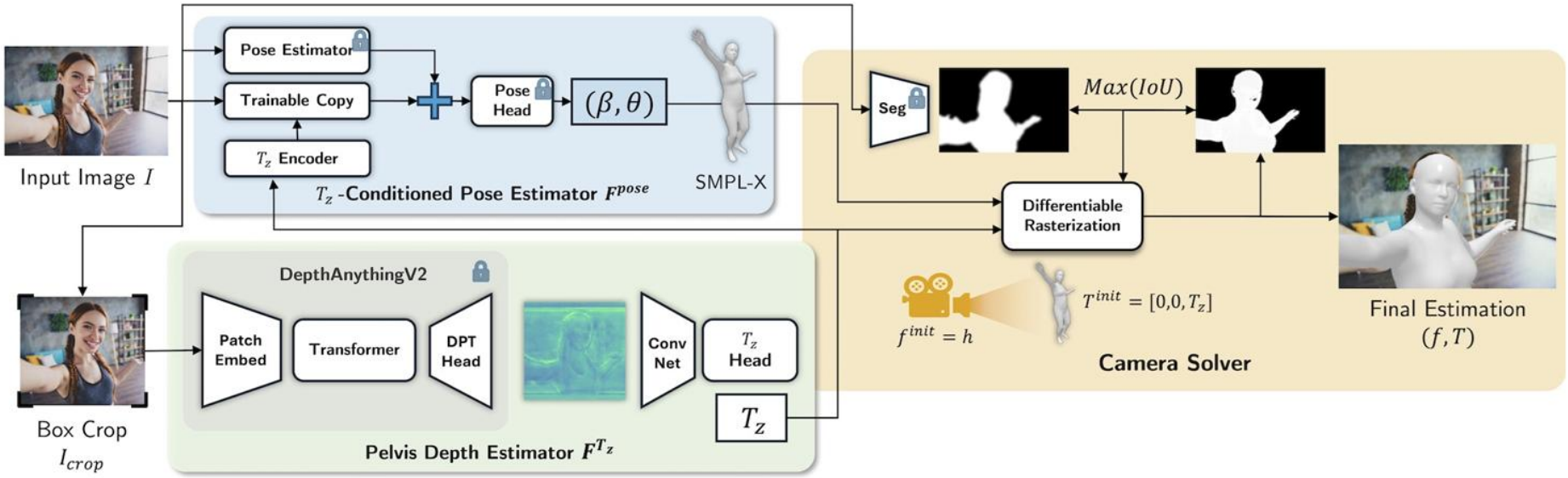




  - 왜곡의 양은 T_z와 강한 상관관계가 있음

Method

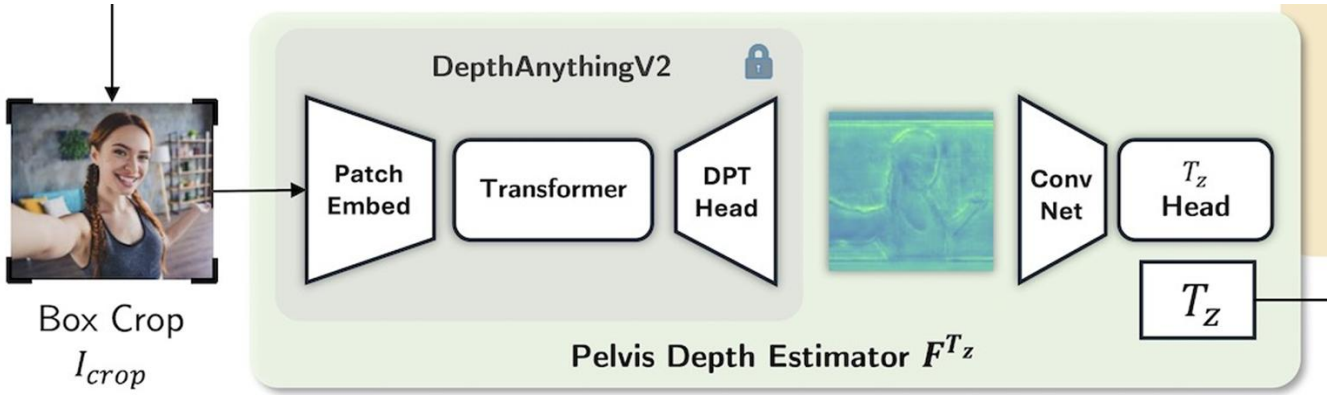
- 3-Step Workflow
  - Predicting Z-Translation T_z
  - T_z-aware Pose Estimation
  - Camera Solver



Method

- Predicting Z-Translation T_z
  - 사람의 Perspective distortion은 카메라 거리 T_z에 의해 결정됨
  - 근거리(<1.2m)에서 왜곡 비선형적으로 증가
  - Pelvis Depth Estimator F_{T_z}
    - Backbone : DepthAnythingV2 (freeze)
    - 입력 : Crop Image
    - 출력 : Pelvis depth T_z
  - Depth Loss

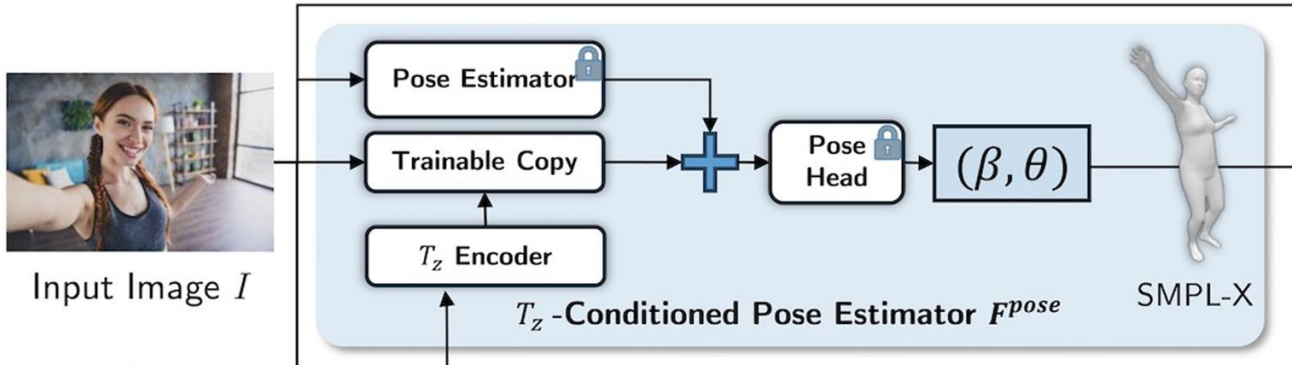
$$L_{depth} = 1/T_z^{GT} \cdot \|T_z - T_z^{GT}\|_1$$



Method

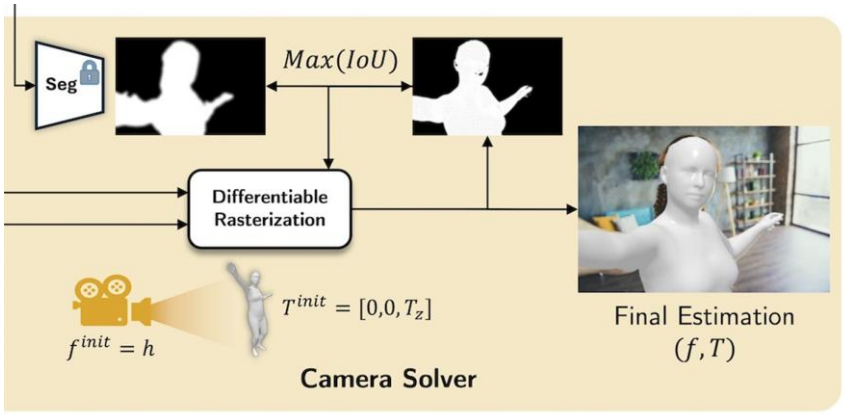
- Tz-aware Pose Estimation
  - AiOS는 카메라와 거리가 먼 사람 위주로 학습하여 근거리 왜곡에 취약
    - AiOS : 단일 이미지로부터 SMPL(-X) 메쉬를 예측하는 HMR 모델
  - Pose Estimator
    - Backbone : Pretrained AiOS (freeze)
    - Trainable copy + Zero-MLP residual
    - 입력 : Image, Predicted T_z
    - 출력 : SMPL-X (β, θ)
  - Pose Network Loss

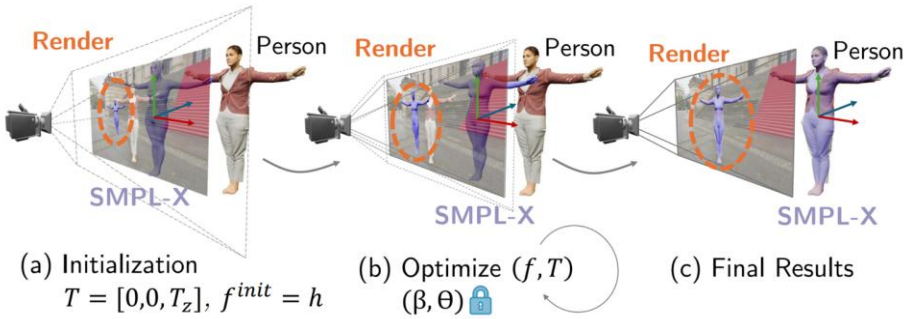
$$L = w_{\text{Shape}}L_{\text{Shape}} + w_{\text{pose}}L_{\text{pose}} + w_{\text{joint}}L_{\text{joint}} + w_{\text{vert}}L_{\text{vert}}$$



Method

- Camera Solver
  - T_z가 고정되면 (f,T_x,T_y)는 단순 alignment 파라미터
  - 초기 설정 : $T=[0,0,T_z]$, f (init) = image height
  - Differentiable Rasterization
    - SMPL-X mesh → binary mask rasterization
    - Segmentation mask 생성
    - IoU 최대화하도록 최적화
  - 최적화 목표
    - Mesh mask vs Segmentation mask (위치, scale 정렬 및 camera parameter 보정)





Experiments

- Quantitative Result

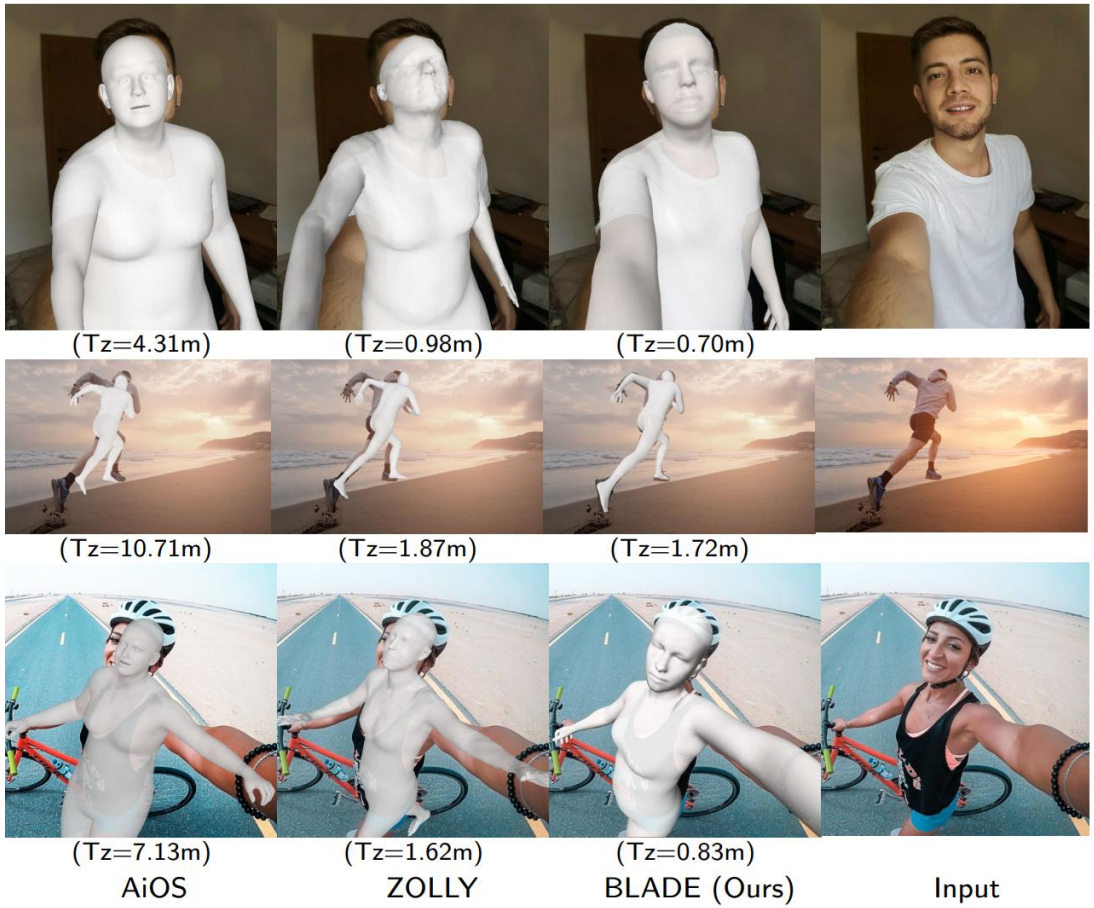
Methods	SPEC-MTP [15] (real-world capture)						PDHUMAN [31] (synthetic)						BEDLAM-CC (synthetic)					
	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑
ZOLLY [31]	0.899	0.394	0.906	1.063	126.7	62.3	0.255	0.355	0.267	0.273	82.0	53.0	0.539	0.634	0.564	0.461	131.8	51.8
SMPLer-X*[29]	0.980	0.450	0.109	1.121	102.6	53.0	2.223	1.030	0.126	0.550	161.2	47.6	2.057	1.172	0.087	1.349	139.9	53.0
TokenHMR*[8]	0.909	0.436	0.095	1.121	124.3	49.7	2.280	1.034	0.068	0.550	156.7	53.0	2.378	1.200	0.096	1.349	136.4	54.2
AiOS*[29]	1.035	0.464	0.121	1.121	110.9	48.7	2.312	1.024	0.149	0.550	183.4	49.5	2.340	1.197	0.111	1.349	143.0	54.6
Ours	0.129	0.114	0.056	0.163	111.9	68.7	**0.106**	**0.176**	**0.043**	**0.216**	**80.5**	**67.3**	0.326	0.305	0.079	0.257	111.6	74.6
Ours (real-world)	**0.127**	**0.112**	**0.044**	**0.159**	**99.6**	**69.5**	0.107	0.178	0.049	0.223	102.6	65.2	**0.325**	**0.305**	**0.076**	**0.212**	**106.8**	**75.0**

<SOTA 메소드와 정량적 비교>

E_{Tz} : z translation 절대 오차
$E_{1/Tz}$: inverse-depth 오차
E_{Txy} : x, y translation L2 오차
E_f : focal length의 상대 오차
PVE : 3D 메쉬 정확도
mIoU : 2D Alignment

Experiments

- Qualitative Result



<SOTA 메소드와 정성적 비교>

Experiments

- Ablation Study
  - Depth backbone
  - Conditioning

	DiNOv2 [27]	Sapiens [12]	DAv2 [33]	Ours
$E_{T_z} \downarrow$	0.300	0.210	0.154	**0.127**

<Depth backbone Ablation study>

	PA-MPJPE↓	MPJPE↓	PVE↓
raw AiOS	62.816	101.577	110.851
ft. AiOS	64.932	113.173	120.582
Ours (T_z cond.)	**56.666**	**94.050**	**99.635**

<Conditioning Ablation study>

Result

- Limitation
  - Single-person image에 대해서만 고려
  - Pinhole camera 이외의 카메라 타입과 lens distortion 고려
  - Segmentation mask가 매우 부정확한 경우 실패
- Conclusion
  - Single image로부터 human mesh recovery와 perspective camera estimation 동시 수행
  - Orthographic camera model 변환 없이 perspective projection parameters 추정 제공

감사합니다