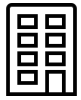
Using Humans as Calibration Objects



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

Presented By
김태우

Contents

- Background
  - Panoptic Studio dataset

- 관련 논문 소개
  - Humans as a Calibration: Dynamic 3D Scene Reconstruction from Unsynchronized and Uncalibrated Videos [ICCV 2025]
  - EasyRet3D: Uncalibrated Multi-view Multi-Human 3D Reconstruction and Tracking [WACV 2025]

1) Joo, Hanbyul, et al. "Panoptic studio: A massively multiview system for social motion capture." *Proceedings of the IEEE international conference on computer vision*. 2015.

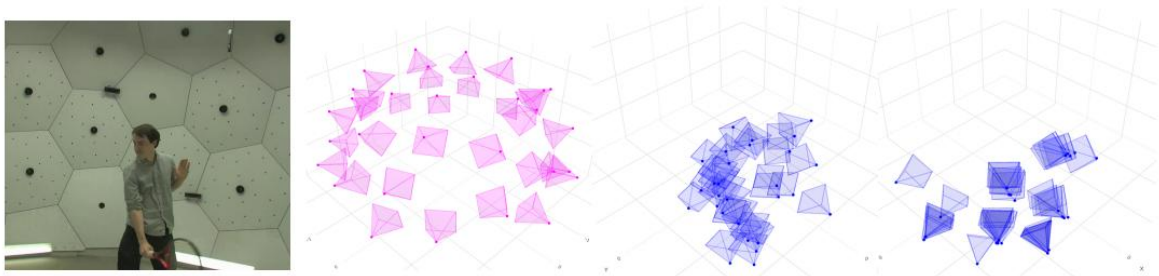
Background

- Panoptic datasets[1] (CMU Panoptic Studio)
  - 데이터 특징
    - 동적 인체 동작 기반 멀티뷰 영상
    - 카메라 구성: 31대 카메라가 반구 형태로 장면을 둘러싸도록 배치
    - 장면 특징: 다양한 인체 동작, 물체와의 상호작용, 시야별 가림(occlusion) 발생
    - 평가 기준: GT(시간 오프셋, 카메라 포즈)를 평가에 사용 가능
  - 학습/평가 뷰 분할
    - 29~30대 카메라로 학습, 1대 카메라로 신규 뷰 합성(novel view) 테스트

<Panoptic datasets>

Background

- Task: 4D Reconstruction in Uncalibrated Dynamic Scenes
  - 비동기(time), 미보정(uncalibrate) 멀티뷰 영상에서 4D 재구성
- 기존 방법의 문제점
  - 기존 4D 기반 방법론은 동기화된 멀티뷰 + 정답 카메라 포즈를 전제
    - 실제 환경 적용성이 낮음
  - 단안 카메라의 비현실적 이동 궤적은 시간 동기화 가정으로 pose 추정
  - DUSt3R/COLMAP: 비슷한 텍스처, 반복 패턴에서 pose 추정에 실패하기 쉬움
- 논문 선정 배경: 현실적 촬영 환경에서 적용 가능한 강건한 방법론 탐색



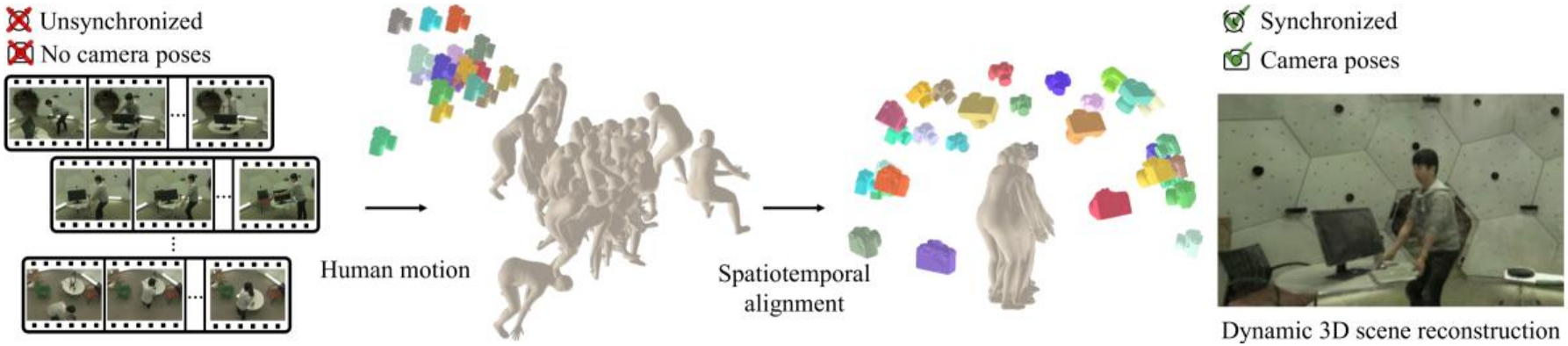
<특징 기반 추출 방법론 실패 사례>

Humans as a Calibration: Dynamic 3D Scene Reconstruction from Unsynchronized and Uncalibrated Videos

Introduction

- Contribution
  - 사람의 포즈,형상 추정을 캘리브레이션 패턴으로 활용
  - 시간 비동기, 포즈 비보정 멀티뷰 영상에서 시간 오프셋과 카메라 포즈를 최적화
  - SOTA 성능 달성, GT(정답 포즈,시간)과 유사한 수준
- Conclusion 요약

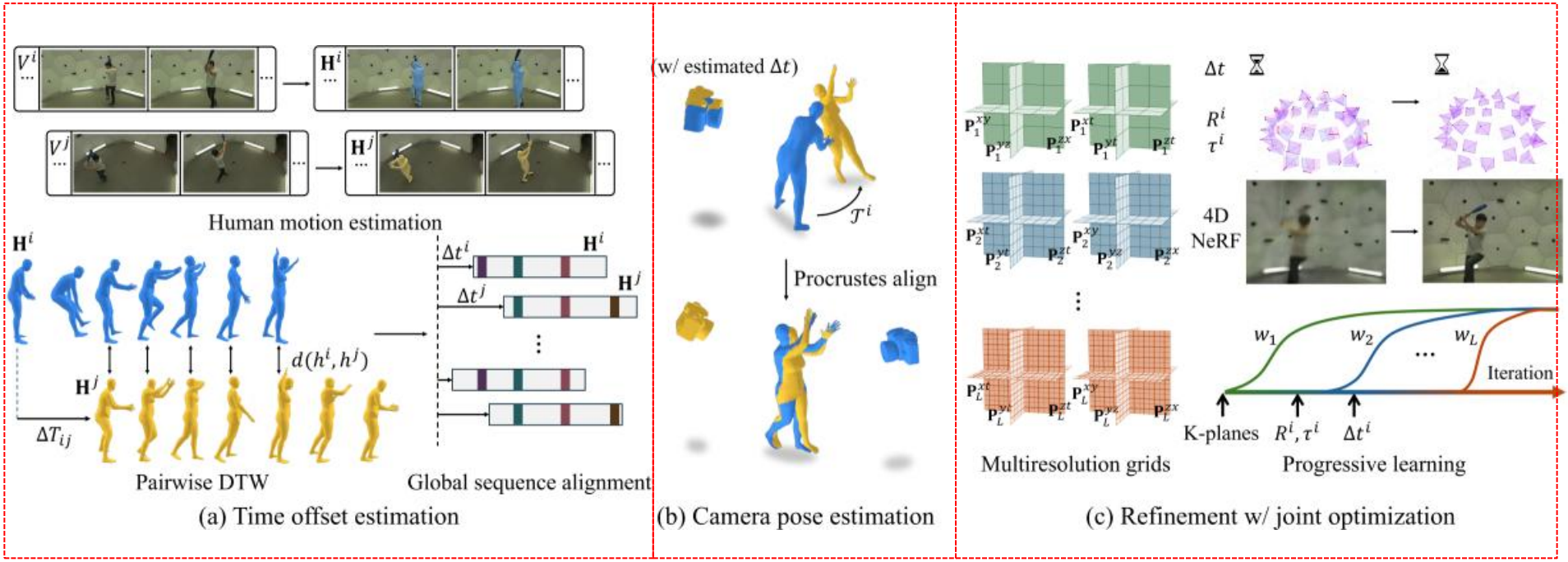
  → 각 프레임 별로 추출한 인간 정보로 Δt, Cam pose의 강건한 초기 추정 가능



<Humans as a Calibration framework>

Method

- Framework overview
  - 총 3파트로 나누어 설명
  - Task 가정
    - 등장하는 인원은 총 1명으로 가정

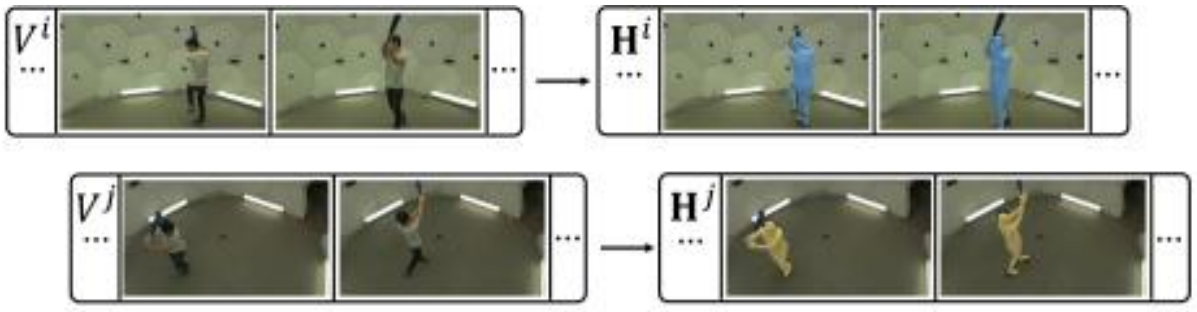


1 2 3

1) Ye, V., Pavlakos, G., Malik, J., & Kanazawa, A. (2023). Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21222-21232).

Method (Part 1)

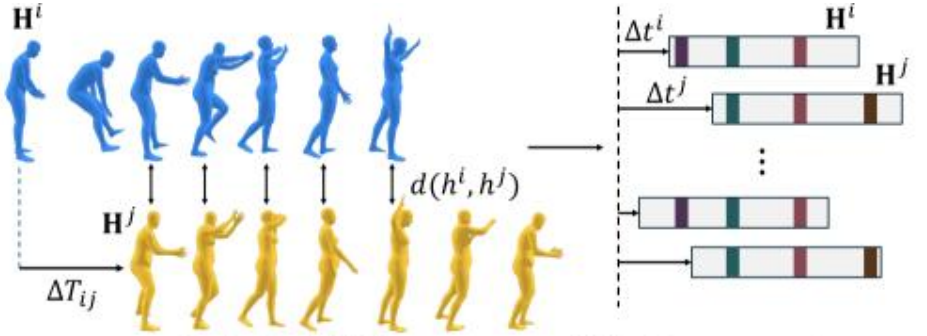
- Data extraction
  - Input: 다중 뷰 비디오 (Uncalibrated)
  - Output: 각 frame 별 Human SMPL (shape, poses)
  - Model: SLAHMR[1] 을 사용해 추정
- Local SMPL coordinates
  - 각 카메라 좌표계마다 SMPL 파라미터를 보유 (shape, poses, trans)



<Frame 별 SMPL 추정>

Method (Part 1)

- Time offset estimation
  - 필요성: 카메라 별 FPS 차이, 여러 다중 뷰 카메라간의 시간(time) 동기화 필요
    - 미 수행시 재구성 성능 저하 발생 (Reprojection 오류)
  - DTW algorism
    - 길이가 다르거나 속도가 다른 두 시퀀스(시계열)를 정렬하는 알고리즘
      - 어느 시점이 서로 대응, 유사도를 측정하는 방법
      - 입력: 각 multi-view 카메라 별 산출된 3D pose trajectory time 시퀀스
  - 문제점 제시
    - 카메라 간의 시간 비동기 오류로, 두 Pose간의 시퀀스가 비슷하지만 시간이 어긋나거나 늘어남/줄어듦이 있을 수 있음



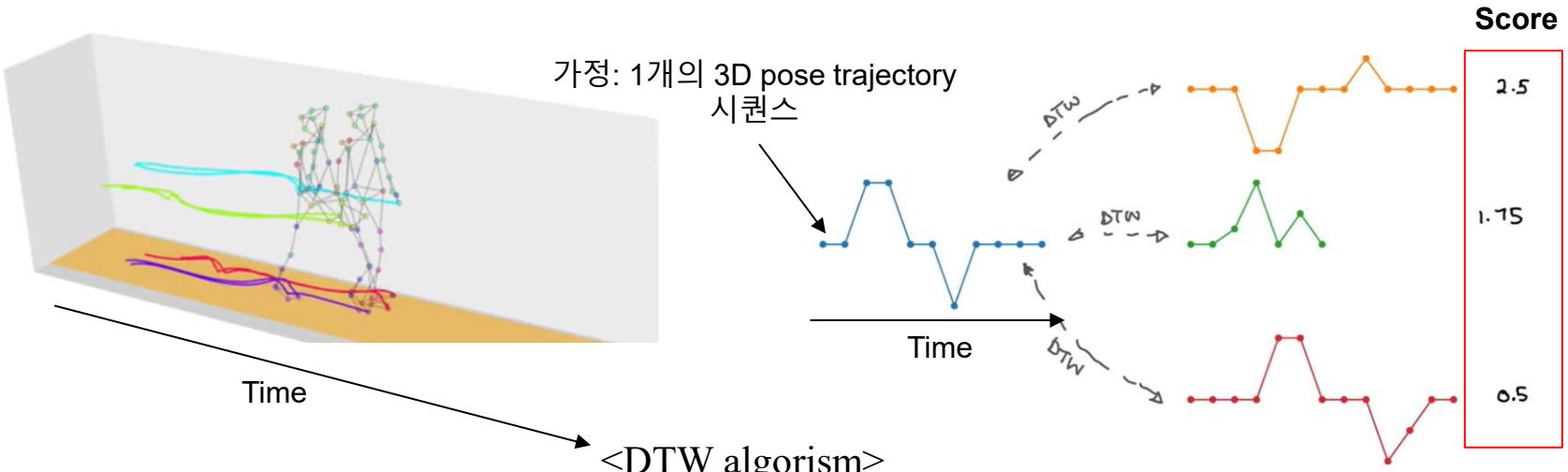
(a) Time offset estimation

Method (Part 1)

- Time offset estimation
  - DTW algorism
    - 같은 인덱스끼리 포인트-투-포인트 비교는 시간 지연/늘어짐을 반영하지 못함
      - 실제로 대응해야 할 피크/패턴을 제대로 비교 불가
    - 두 신호의 지점들을 유연하게 매칭하여, 한 지점이 다른 쪽의 여러 지점과도 대응 가능
  - 정렬 비용: DTW는 정렬의 총 비용을 정의, 가장 작은 score값이 제일 유사

$$d(h_{t_1}^i, h_{t_2}^j) = \|\mathbf{J}_{\text{canon},t_1}^i - \mathbf{J}_{\text{canon},t_2}^j\|_2, \qquad \Delta t = \text{GLOBAL ALIGN}(C, \Delta T),$$

Distance 정의 — 카메라 별 time offset 산출



<DTW algorism>

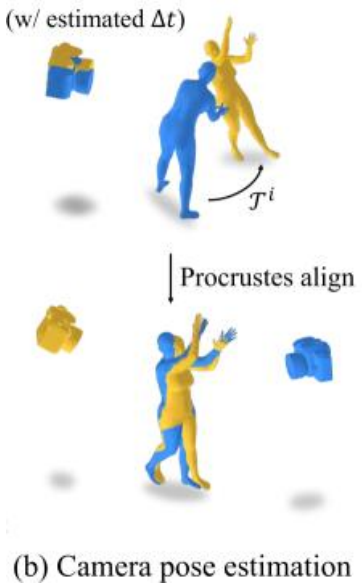
테스트 보행 신호와 3개 후보 시퀀스 간 DTW 비용 비교

Method (Part 2)

- Camera pose estimation
  - 카메라 포즈 추정 (Procrustes 기반 정합)
    - 핵심 아이디어
      - 사람의 3D 관절을 캘리브레이션으로 사용
      - 해당 변환을 카메라 포즈에도 동일하게 적용하여 전역 좌표로 변환
      - 앞서 계산한 시간 Δt 변환 기준 3D pose 기반 정합
    - Procrustes 기반 정합 사용
      - 앵커 카메라 α를 무작위로 선택 (α∈[1,N])
      - 모든 i(카메라 index)≠α 에 대해, 앵커와의 정합을 수행
      - 해당 변환을 카메라 포즈에도 동일하게 적용(함께 이동)

$$s_i, s_\alpha, \mathbf{t}_i, \mathbf{t}_\alpha, R = \text{PROCRUSTES}((\mathbf{J}^i_{\text{global},t+\Delta t^i}; t), (\mathbf{J}^\alpha_{\text{global},t+\Delta t^\alpha}; t)).$$

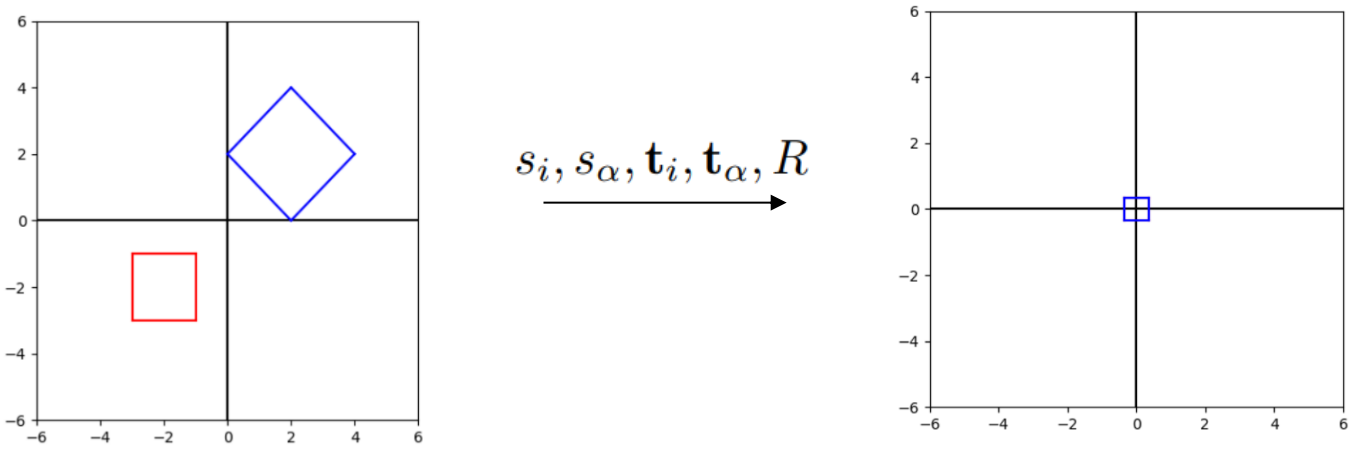
<Procrustes 기반 정합>



(b) Camera pose estimation

Method (Part 2)

- Procrustes 기반 분석
  - 목적: 형상 유사도를 비교
  - 방법
    - 한 쪽을 이동(Translation), 스케일(Uniform scaling), 회전(Rotation) 하여, 다른 쪽에 최적 정렬하는 방법
    - 절차는 3단계로 구성 (Translation, scaling, Rotation)



< 매칭 문제: Procrustes 기반 정합>

Method (Part 2)

- Procrustes 기반 분석
  - 목적: 형상 유사도를 비교
  - 방법
    - 1. Translation을 이용하여 두 도형의 중심점을 맞춤
      - 중심의 정의는 모든 Shape 좌표에 대한 평균으로 계산
    - 2. Normalization을 통해 두 도형의 크기(Scale)를 맞춤
    - 3. Rotation
      - 3-1. 기준 객체 설정: 한 객체를 기준 객체로 설정하고 다른 객체를 회전
      - 3-2. 최적 회전 각도 계산: 두 객체의 점 사이 거리 제곱합(SSD)을 최소화 회전 각도

Uniform scaling [edit]

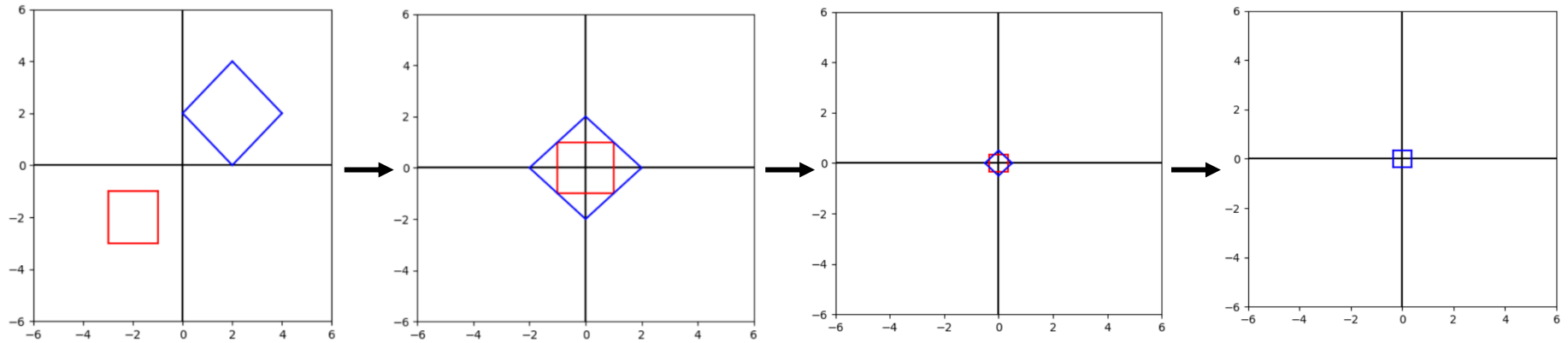
Likewise, the scale component can be removed by scaling the object so that the root mean square distance (*RMSD*) from the points to the translated origin is 1. This RMSD is a statistical measure of the object's **scale** or **size**:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (y_1 - \bar{y})^2 + \cdots}{k}}$$

The scale becomes 1 when the point coordinates are divided by the object's initial scale:

$$((x_1 - \bar{x})/s, (y_1 - \bar{y})/s).$$

Uniform scaling



Translation (대칭 이동) Uniform scaling Rotation

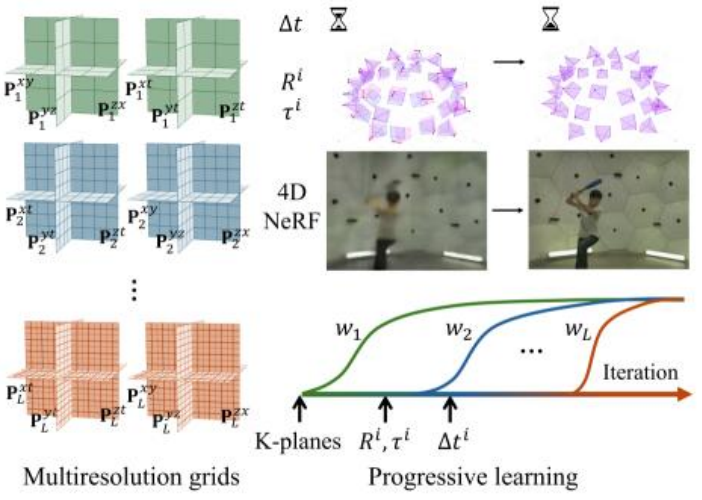
Method (Part 3)

- 4D Reconstruction with camera refinement
  - 목적: 시간 오프셋(Δt) 과 카메라 포즈(R,T)를 동적 NeRF 학습으로 최적화 및 검증
    - 표현 방법
      - K-Planes 기반 4D NeRF
      - 6개 평면의 다중 해상도 특징 gird 사용 (Hex-plane)
  - Progressive learning
    - 커리큘럼 학습으로 파라미터 점진적 학습
  - 시간 오프셋 (Δt) 고려 랜더링
    - Δt 를 더한 time의 feature 추출하여 랜더링
    - GT ←→ Rendered image 사이 L2 pixel loss

$$\hat{I} = \sum_{k=1}^{N} T_k (1 - e^{\sigma(\mathbf{x}_k, t+\Delta t^i)\delta_k}) \mathbf{c}((\mathbf{x}_k, t+\Delta t^i), \mathbf{d}_t^i),$$

$$\mathcal{L} = \sum_{i,t,\mathbf{r}} \| I_t^i(\mathbf{r}) - \hat{I}(\mathbf{r}_t^i, t+\Delta t^i) \|_2,$$

<Loss 를 통한, (R,T,Δt) 최적화>



(c) Refinement w/ joint optimization

1) Xu, Z., Peng, S., Lin, H., He, G., Sun, J., Shen, Y., ... & Zhou, X. (2023). 4k4d: Real-time 4d view synthesis at 4k resolution. 2024 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 20029-20040).

Experiments

- Mobile-Stage dataset[1)]
  - 구성: 스마트폰 멀티뷰로 촬영된 3명의 댄서(정면 + 측면 뷰)
  - 데이터 특징
    - 일부 시점에서 심한 가림(occlusion) 발생
  - 학습/평가 세팅
    - 훈련 20대 카메라, 평가 1대 카메라 사용
  - 가림과 시야 불일치가 많아 강건한 멀티뷰/동적 장면 재구성 성능 검증에 적합

<Mobile-Stage dataset>

1) Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., ... & Tang, S. (2022, October). Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision* (pp. 180-200). Cham: Springer Nature Switzerland.

Experiments

- EgoBody dataset[1)]
  - 장면 구성: 2인의 인간 상호작용
  - 카메라 세팅
    - 정적 Kinect: (S22-S21-02: 4대, S32-S31-01, S32-S31-02: 5대)
    - 이동형 HoloLens2: 머리 착용식 카메라 각 scene 별 1대
    - 비고: 데이터 품질 이슈 존재하는 데이터 세트
      - HoloLens2 영상에 결손 프레임 존재, 강한 모션 블러 프레임 다수
      - 저자는 선형 보간으로 보완하여 해결
  - 시퀀스 길이: 멀티뷰 학습 비디오 각 200프레임

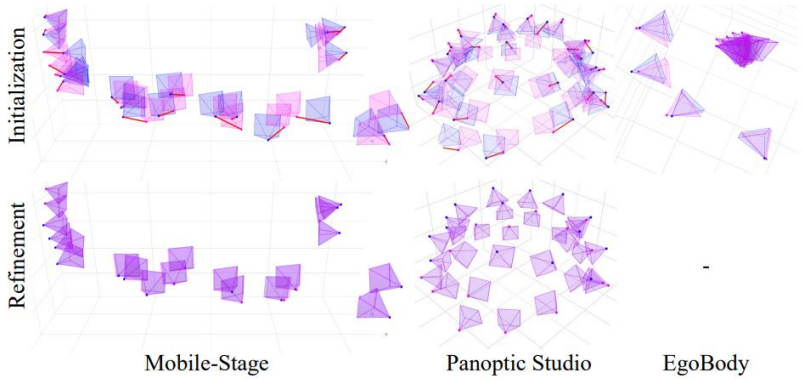
<EgoBody dataset>

Experiments

- 정량적 수치 결과 (Camera, time refinement)
  - 결론: 좋은 초기값을 제공, 이를 보여주는 것이 목적
    - Egobody 데이터 세트의 refine 은 따로 수치로 다루지 않음

Dataset	Scene	Rotation (°)		Trans.		Δt (frames)		
		Init	Refine	Init	Refine	Data	Init	Refine
Panoptic Studio	BASEBALL	5.30	0.328	22.6 cm	0.16 cm	29.04	0.700	0.025
	OFFICE1	3.88	0.420	19.7 cm	0.17 cm	33.25	3.448	0.031
	OFFICE2	8.89	0.660	38.0 cm	0.37 cm	29.65	1.300	0.029
	OFFICE3	3.95	0.363	19.2 cm	0.17 cm	32.93	0.800	0.026
	TENNIS	5.29	0.290	25.8 cm	0.22 cm	28.33	0.467	0.028
	Average	5.46	0.412	25.1 cm	0.22 cm	30.64	1.343	0.028
Mobile-Stage	DANCE	5.65	1.707	0.053	0.002	24.76	0.455	0.214
EgoBody	S22-S21-02	12.7	-	0.016	-	21.94	7.200	-
	S32-S31-01	3.68	-	0.025	-	33.30	0.333	-
	S32-S31-02	10.8	-	0.034	-	8.333	1.667	-
	Average	9.07	-	0.025	-	21.19	3.067	-

<Camera, time refinement 결과>



<Cam pose, Red: GT, Blue: Estimation>

Experiments

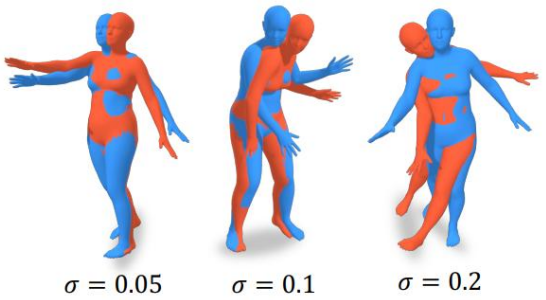
- 노이즈 상황 가정 실험 (robustness 평가)
  - 노이즈 종류
    - 카메라 별 Frame rate 가 다른 가정 상황 실험
      - 조건: FPS=24 for five videos and FPS=30 for others
    - Video degradation (비디오 열화)
      - 조건: 비디오 열화: 감마($\gamma \in [0.35, 2.1]$), 휘도/색 노이즈 추가
    - SMPL 노이즈: β, Θ에 가우시안 노이즈(σ=0.01~0.2)
  - 결론: 인간 동작은 강력, 강건한 표현으로, time 불일치, uncalibrated 환경에서도 신뢰도 높은 초기화

TENNIS scene	Rotation (°)		Trans. (cm)		Δt (frames)	
	Init	Refine	Init	Refine	Init	Refine
L2 norm, β, Θ	5.382	0.656	26.130	0.261	2.100	0.030
SMPL noise, $\sigma = 0.01$	5.266	0.285	25.844	0.296	1.633	0.027
SMPL noise, $\sigma = 0.02$	5.307	1.240	25.825	0.276	0.533	0.028
SMPL noise, $\sigma = 0.05$	5.310	0.357	25.861	0.277	0.833	0.029
SMPL noise, $\sigma = 0.1$	5.241	0.518	25.383	0.257	2.100	0.030
SMPL noise, $\sigma = 0.2$	5.404	0.934	26.831	0.217	2.367	0.024
Degraded video	6.496	-	32.374	-	1.133	-
Mixed FPS	7.933	-	38.58	-	1.200	-
Default setup	5.293	0.290	25.827	0.217	0.467	0.028

<노이즈에 따른 metric 비교>

(a) Video degradation

<Video degradation 노이즈>



(b) Noised human motion

<SMPL 노이즈>

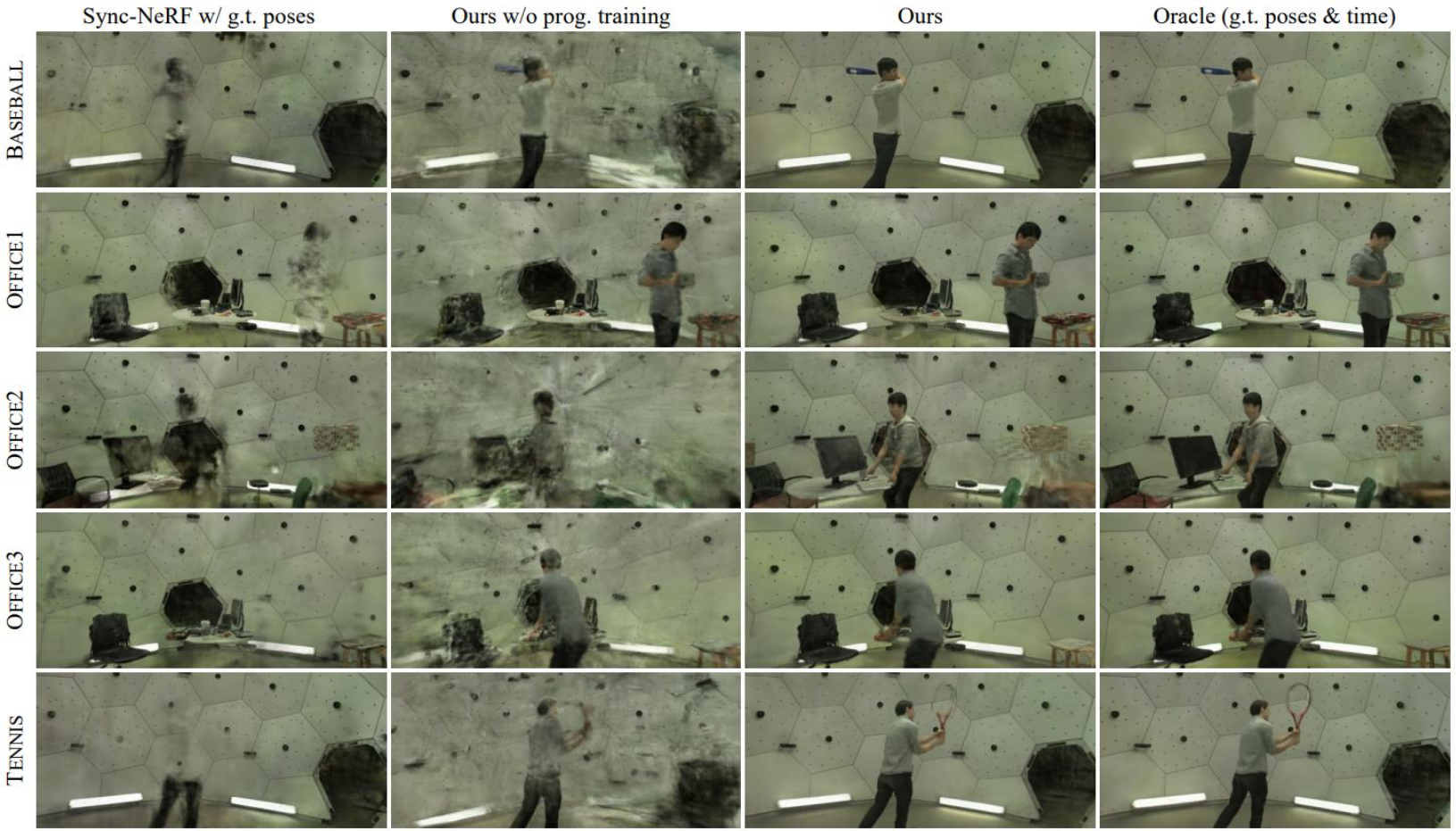
Experiments

- 정량적 수치 결과
  - PSNR / SSIM / LPIPS 비교
  - 입력 영상이 비동기이므로, 모든 학습 영상에 공통으로 겹치는 부분으로 평가
- 비교 모델
  - Oracle: k-planes 기반 방법, GT cam pose & GT time 제공
  - Sync-NeRF[1]: Time refine 가능, 카메라 최적화 불가 → 비교를 위해 GT cam pose
- 결론
  - Cam pose/time 값 없는 환경에서도 Oracle 수준에 근접
  - 일부 장면에서는 PSNR 상회

Dataset	Scene	Sync-NeRF w/ GT pose			Ours w/o prog. training			Ours			Oracle (GT pose & time)		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Panoptic Studio	BASEBALL	21.17	0.833	0.189	19.23	0.577	0.412	27.20	0.919	0.072	26.80	0.925	0.065
	OFFICE1	20.62	0.809	0.196	22.33	0.678	0.334	26.65	0.855	0.125	27.00	0.905	0.079
	OFFICE2	21.14	0.782	0.195	18.39	0.511	0.544	24.43	0.820	0.155	26.58	0.891	0.091
	OFFICE3	20.94	0.826	0.178	21.06	0.614	0.383	27.51	0.893	0.093	28.23	0.912	0.077
	TENNIS	22.09	0.851	0.168	17.29	0.531	0.552	26.94	0.881	0.113	27.22	0.916	0.080
	Average	21.19	0.820	0.185	19.66	0.582	0.445	26.55	0.874	0.112	27.16	0.910	0.078
Mobile-Stage	DANCE	17.26	0.410	0.284	13.72	0.233	0.510	23.27	0.816	0.089	22.98	0.806	0.093

Experiments

- 정성적 결과



<Panoptic studio 정성적 결과>

EasyRet3D: Uncalibrated Multi-view Multi-Human 3D Reconstruction and Tracking

Background

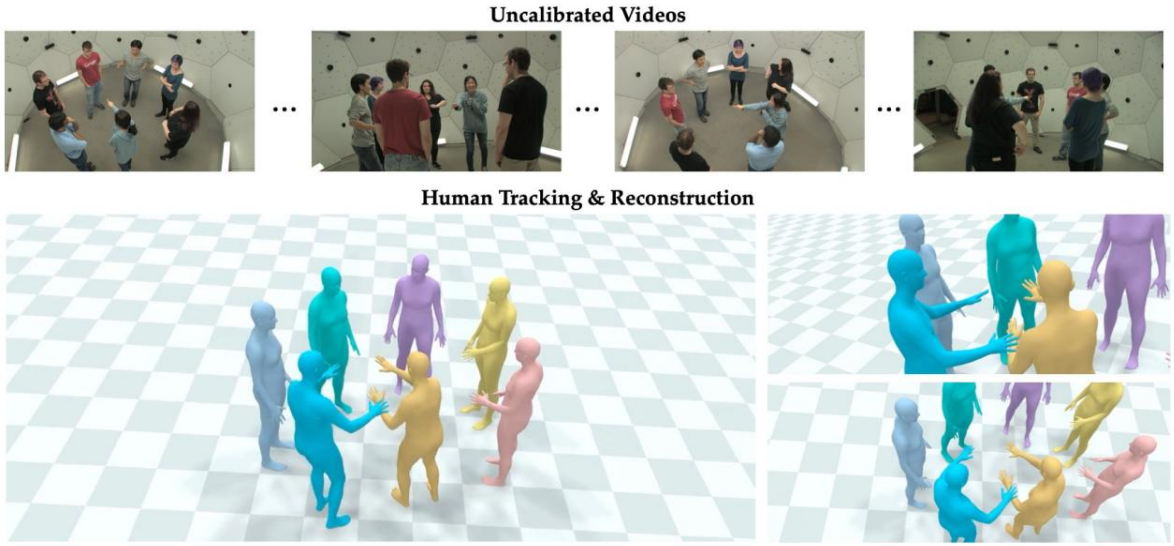
- 기존 문제점
  - Human 3D Reconstruction 성능이 카메라 추정(내,외부 파라미터) 품질에 과도하게 의존
- 선정 배경
  - 하나의 뷰만을 사용해, SMPL 추정시 가려질 경우 추정 불가
  - 실상황은 가려짐 다수 발생
    - 멀티뷰를 함께 사용해, 가려짐에 강건하고, 다중 인원에 대한 개별적 처리 가능

<Multi-view, Multi-person data>

Introduction

- Contribution
  - 다중 카메라로 촬영된 영상에서 여러 사람의 3D 포즈를 추정
  - 인물의 3D 포즈 정보를 이용, 카메라 파라미터를 자동으로 추정
  - 적응형 스티칭 모듈 설계, 신뢰도에 따라 병합하여 정확한 global SMPL 모델 출력
- Conclusion 요약

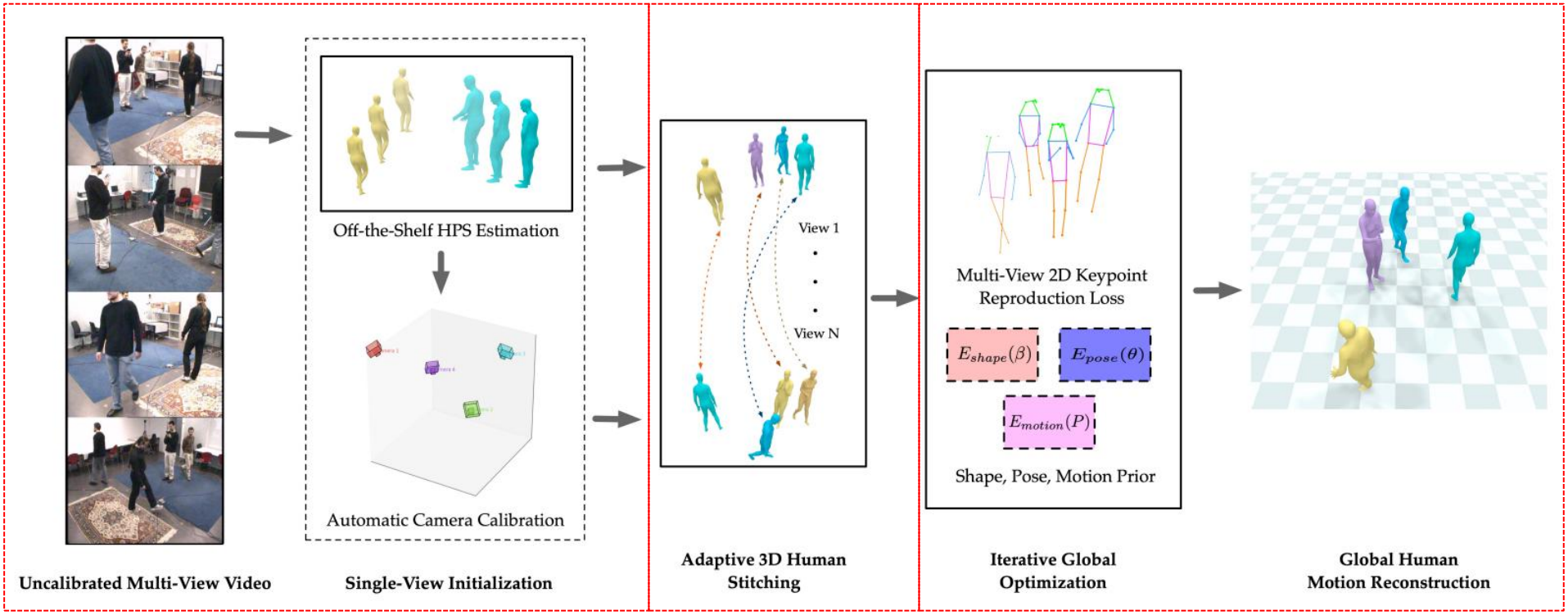
  → 수동 카메라 보정 없이, 다중 뷰 영상에서 SOTA 성능 달성



<EasyRet3D>

Method

- Framework overview
  - 총 3파트로 나누어 설명

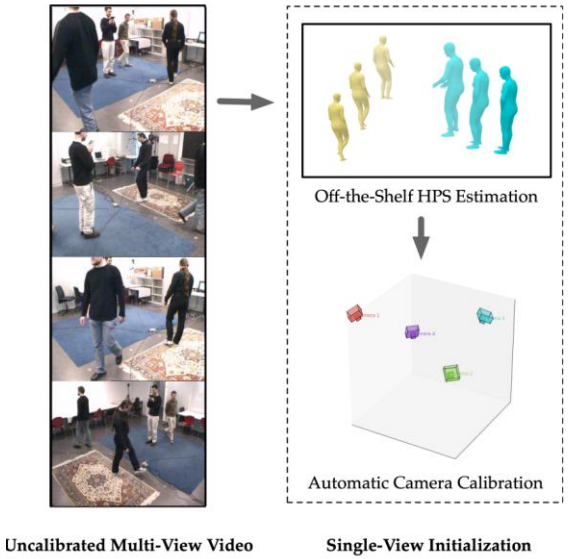


1 2 3

1) Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., & Malik, J. (2023). Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 14783-14794).
2) Rajasegaran, J., Pavlakos, G., Kanazawa, A., & Malik, J. (2022). Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2740-2749).

Method (Part 1)

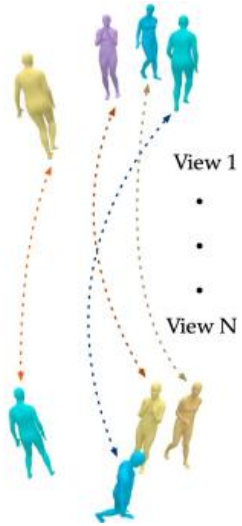
- Data extraction
  - Input: 다중 뷰 비디오 (Uncalibrated)
  - Output: 각 frame 별 Human SMPL (shape, poses)
  - Model: HMR2.0[1] 기반 PHALPS[2] 모델 사용
- Local SMPL coordinates
  - 각 카메라 좌표계마다 SMPL 파라미터를 보유 (shape, poses, trans)



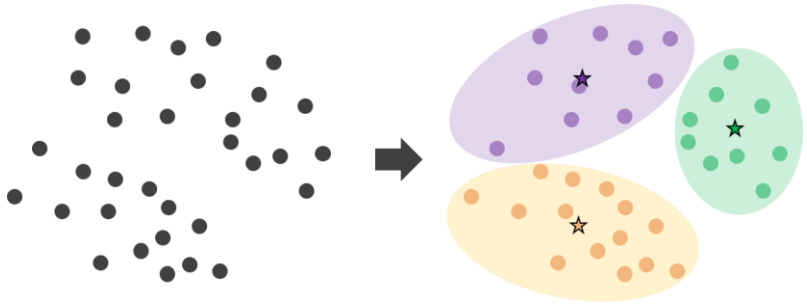
1) Rajasegaran, J., Pavlakos, G., Kanazawa, A., & Malik, J. (2022). Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2740-2749).

Method (Part 1)

- 동일 인물 추적 (Cross-View Matching)
  - 필요성: 다중 카메라의 각 뷰의 추정되는 SMPL 간의 매칭이 없음
- 매칭 방법
  - Human feature 추출
    - Human cropped image → PHALPS[1] → 3D appearance embedding features
  - Feature 간 클러스터링
    - 각 point 간의 L2 거리합을 최소화하는 방향으로 클러스터링하여 구분



<Multi-view SMPL 간 매칭 필요>


<Feature 간 클러스터링>

1) https://en.wikipedia.org/wiki/Random_sample_consensus

Method (Part 2)

- Camera initialization
  - 목적: 3D Human pose를 사용해, 카메라 파라미터(R, T)의 초기값을 추정
  - 방법
    - Input: Local 3D Poses (multi-view), 2D keypoints (multi-view)
    - 1. Multi-view 2D keypoint 중 가장 신뢰도가 높은 3D Pose 선택 (Global 로 가정)
    - 2. Cross-View Matching 으로 동일 인물의 3D pose 을 기준으로 선정
    - 3. 해당 view의 3D pose 를 기준으로, 나머지 view 의 2D keypoint 간의 PnP 매칭
      - 카메라 변환 행렬과 이동 벡터를 RANSAC[1] 기반 PnP로 반복적으로 계산
    - 가정: 카메라는 모두 고정되어 있으며, 모두 같은 내부 파라미터를 공유

→ 초기값 산출 이후 반복 최적화 단계에서 파라미터를 추가 조정

$$P_{C_v,z} = R_W^{C_v} P_{W,z} + T^{C_v} \quad z = 1, \ldots, Z$$

<Global to Local 좌표계 변환 공식>

$P_{W,z}$: 월드(공통) 좌표계에 정의된 3D 포인트 (예: 관절 위치)

$R_W^{C_v}$: 월드 좌표계에서 카메라 v 좌표계로의 **회전 행렬**

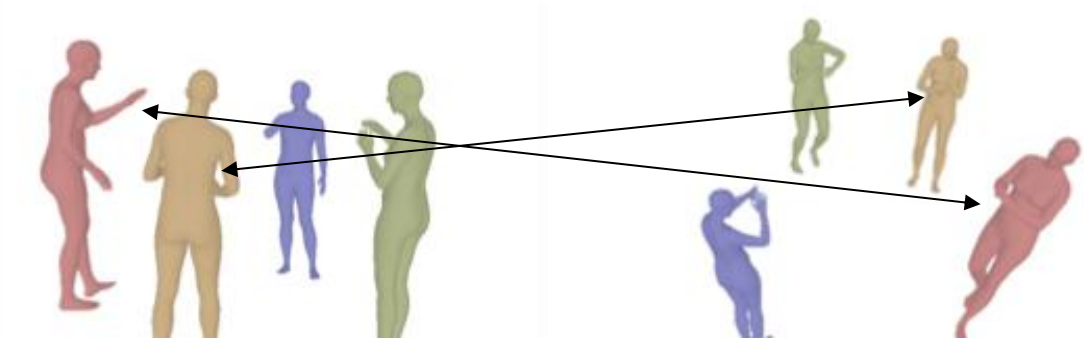
T^{C_v} : 월드 좌표계에서 카메라 v 좌표계로의 **이동 벡터 (translation)**

$P_{C_v,z}$: 카메라 v 좌표계에서의 해당 3D 포인트의 위치

z : 각 3D 포인트(예: 관절 인덱스), 총 Z개

Method (Part 2)

- Adaptive 3D Human Stitching Module
  - 필요성: 특정 뷰에서 가림, 조명 변화 현상으로 특정 SMPL 예측 실패
    - 신뢰성 있는 Global SMPL 구축 필요
  - 방법: 각 Local SMPL 간의 신뢰도 기반 가중합
    - 각 view 별, 2D keypoint 의 신뢰도를 기반으로, 추정이 잘된 SMPL 의 파라미터를 가져옴
    - 예: frame 1의 SMPL 정보 → 다수의 가림이 존재하는 frame 2: SMPL 정보 가중합



<View 간의 SMPL 파라미터 차이 존재>

Method (Part 3)

- 최적화 단계
  - Iterative Global Optimization Module
    - 총 3단계로 구성
    - 각 단계별로 목표하는 최적화 상이
      - Stage 1: 2D Reprojection loss만 사용
      - Stage 2: Priors 추가 (Human shape, pose)
      - Stage 3: Motion/Environment 제약 추가
  - Stage 1.
    - Loss: L_{J2D}
    - Reprojection error: Global SMPL 의 2D Keypoint 재투영 오차
    - 최적화 대상: 카메라 파라미터, SMPL (pose)

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{J2D}}.$$

$$\mathcal{L}_{\text{J2D}} = \sum_{v=1}^{V}\sum_{i=1}^{N}\sum_{t=1}^{T}\zeta^{i,v}\rho(\Pi_k({}^{v}R_t\,{}^{w}J_t^i + T_t^v) - x_t^{i,v}).$$

<Reprojection error Loss>

Method (Part 3)

- 최적화 단계
  - Iterative Global Optimization Module
  - Stage 2.
    - Priors 추가 (Human shape β , pose Θ)
    - Loss: $\lambda_{J2D}L_{J2D} + \lambda_{\beta}L_{\beta} + \lambda_{\Theta}L_{\Theta} + \lambda_{smooth}\, L_{smooth}$
  - Smoothing Loss
    - Temporal consistency 강화 (프레임 간 급격한 변화 억제)

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{J2D}}\mathcal{L}_{\text{J2D}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\Theta}\mathcal{L}_{\Theta} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}.$$

$$L_{\beta} = \sum_{i=1}^{N} \|\boldsymbol{\beta}_i\|_2^2, \qquad L_{\Theta} = \sum_{i=1}^{N}\sum_{t=1}^{T} \|\iota_i^t\|_2^2$$

Vposer 입력으로 인코딩 파라미터

<Reprojection error + Human prior Loss>

Method (Part 3)

- 최적화 단계
  - Iterative Global Optimization Module
  - Stage 3.
    - Motion priors, 환경 제약 loss 추가
    - Loss: $\lambda_{J2D}L_{J2D} + \lambda_{\beta}L_{\beta} + \lambda_{\Theta}L_{\Theta} + \lambda_{smooth}L_{smooth} + \lambda_{motion}L_{motion} + \lambda_{env}L_{env}$
  - Motion Loss
    - NeMF 기반 분포 모델링
    - 기존 latent 상 분포에서 잘못된 motion 이 나오지 않도록 제약
  - Environment (env) Loss
    - 발이 지면이 떨어져 있을 경우, 페널티 부여 (Skating 방지)

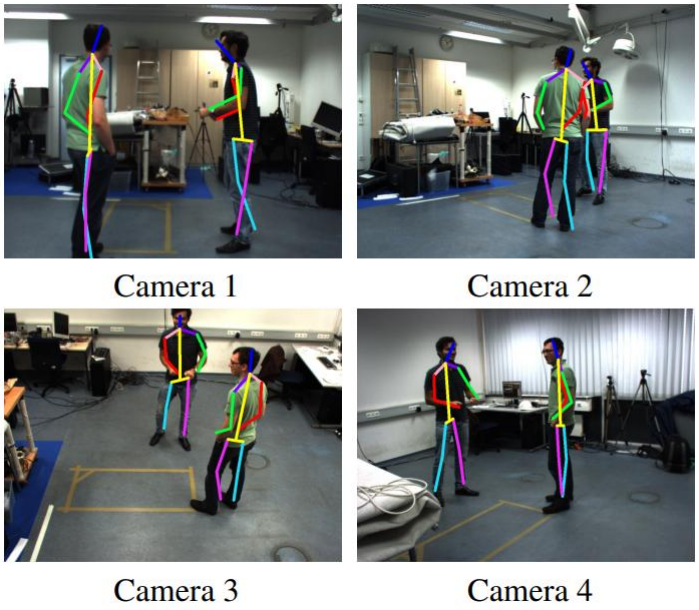
$$\mathcal{L}_{stage3} = \lambda_{J2D}\mathcal{L}_{J2d} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\Theta}\mathcal{L}_{\Theta} + \lambda_{motion}\mathcal{L}_{motion} + \lambda_{env}\mathcal{L}_{env}.$$

<Final Loss>

1) Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2014). 3D pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1669-1676).

Experiments

- Shelf dataset[1)]
  - 데이터 특징
    - 카메라 구성: 5대
    - 특성: 4명이 서로 상호작용하며 선반을 분해하는 복잡한 시나리오
    - 다중 인원 + 근접 상호작용 → 가림(occlusion) 접촉 발생



<Shelf dataset>

1) Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, *36*(7), 1325-1339.

Experiments

- Human3.6M[1)]
  - 데이터 특징
    - 동기화·보정된 4대 카메라 사용
    - 각 포즈에 이미지,비디오와 3D 관절 GT 제공
      - 3.6M(360만) 3D 인체 포즈 GT 제공
    - 특징: 통제된 실내에서 촬영, 3D 포즈 추정 표준 벤치마크

<Human3.6M dataset>

Experiments

- 정량적 수치 결과
  - Metric: MOTA, IDF1, ID Switch
    - MOTA: 탐지 실패·오탐·ID 스위치를 모두 반영한 종합 정확도
    - IDF1: 같은 사람을 전체 scene에 대해, 동일한 ID로 잘 추적했는지 일관성 정확도
    - ID Switch: 같은 GT 인물을 추적하는 도중, 추적기의 ID가 바뀐 횟수
    - PA-MPJPE: 스케일을 고려하지 않은 형태(포즈) 정확도
  - Train Free: Test data 에 대해, 추가 fine-tuning 불필요
  - Calib Free: 수동적으로 파라미터를 조정할 필요가 없음

Dataset	Method	Type	Calib. Free	Train. Free	MOTA↑	IDF1↑	ID Switch↓
Panoptic	Snipper [50]	Mono	✗	✗	93.4	85.5	-
	VoxelTrack [48]	Multi	✗	✗	98.4	98.6	**0**
	TEMPO [7]	Multi	✗	✗	98.4	93.6	-
	Ours	Multi	✓	✓	**98.9**	**98.8**	**0**
Shelf	Yang et al. [42]	Multi	✗	✗	94.6	-	-
	VoxelTrack [48]	Multi	✗	✗	94.4	97.2	**0**
	Ours	Multi	✓	✓	**97.2**	**99.0**	**0**

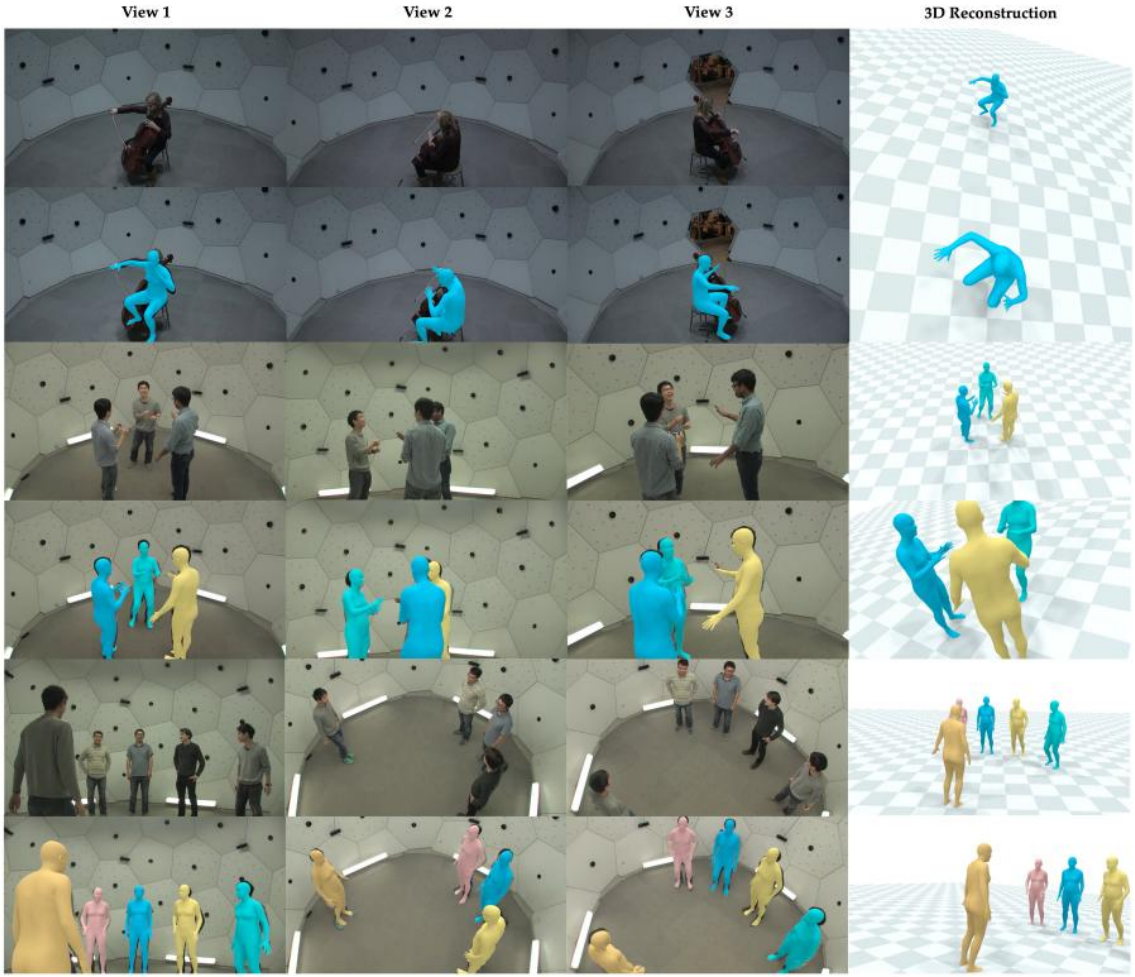
<Panoptic, shelf dataset 평가 결과>

Shelf	Type	Calib. Free	Train. Free	Actor 1 ↑	Actor 2 ↑	Actor 3 ↑	Avg. ↑	PA-MPJPE (mm) ↓
VoxelTrack [48]	Multi	✗	✗	98.6	94.9	97.7	97.07	-
Wu et al. [38]	Multi	✗	✗	99.3	96.5	97.3	97.70	-
VoxelPose [37]	Multi	✗	✗	99.2	95.1	97.8	97.37	-
VTP [5]	Multi	✗	✗	99.5	96.2	97.6	97.77	-
Chen et al. [4]	Multi	✗	✓	99.6	93.2	97.5	96.77	47.6
Huang et al. [13]	Multi	✓	✓	**99.8**	96.5	97.6	97.97	45.9
TEMPO [7]	Multi	✗	✗	99.0	96.3	**98.2**	98.0	43.1
Ours	Multi	✓	✓	**99.8**	**97.2**	97.9	**98.30**	**41.7**

<Human3.6M dataset 평가 결과>

Experiments

- 정성적 결과



<정성적 결과>