『2026 동계 세미나』

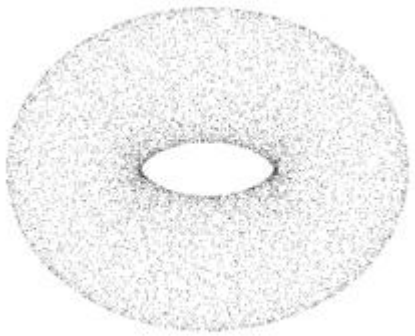
A Unified Feedforward Framework for Scene and Human Understanding

Pusan National University
Dept. of Electronics Engineering
Computer Vision & Signal Processing Lab.

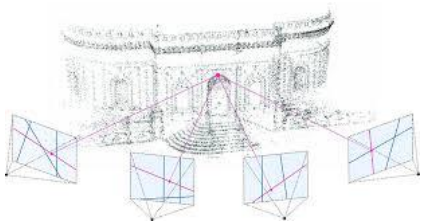
Presented By
최지훈

Background

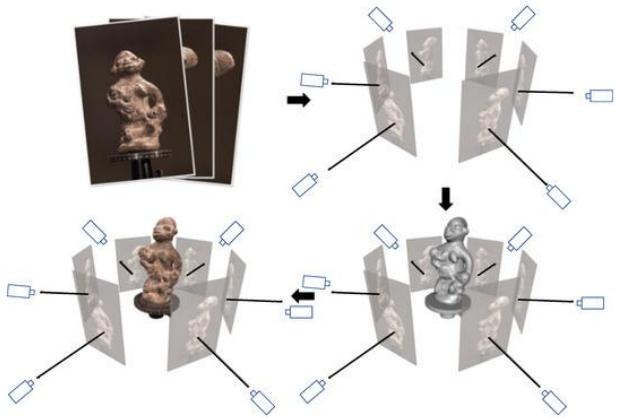
- Point Cloud
  - 3D 공간 상의 점들의 집합으로, 가장 기본적인 3D 표현 방식
    - 여러 장의 사진에서 같은 특징점을 찾아 연결하는 SfM(Structure from Motion), 여러 사진을 대조하여 물체의 위치가 조금씩 다른 점을 이용해 복원하는 MVS(Multi View Stereo) 기법 등이 존재함



< 3D Point Cloud >



< SfM >



< MVS >

3D Reconstruction

- Task
  - Depth Map
    - 이미지의 각 픽셀이 카메라로부터 얼마나 떨어져 있는지를 나타낸 사진
    - 카메라의 위치를 알면 이를 통해 각 픽셀의 위치를 구해 Point Cloud를 구할 수도 있다



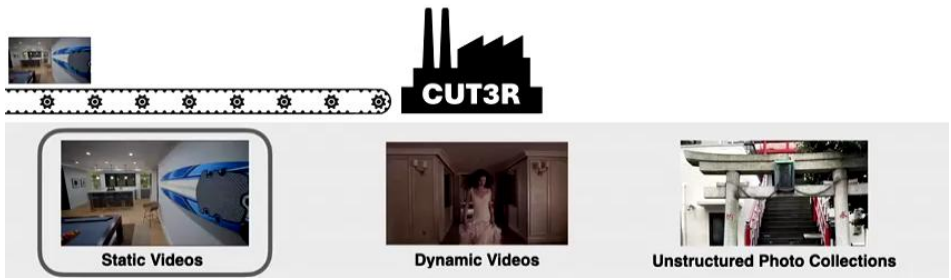
< Reference >

< Depth Map >

Continuous 3D Perception Model with Persistent State
[CVPR 25] Oral

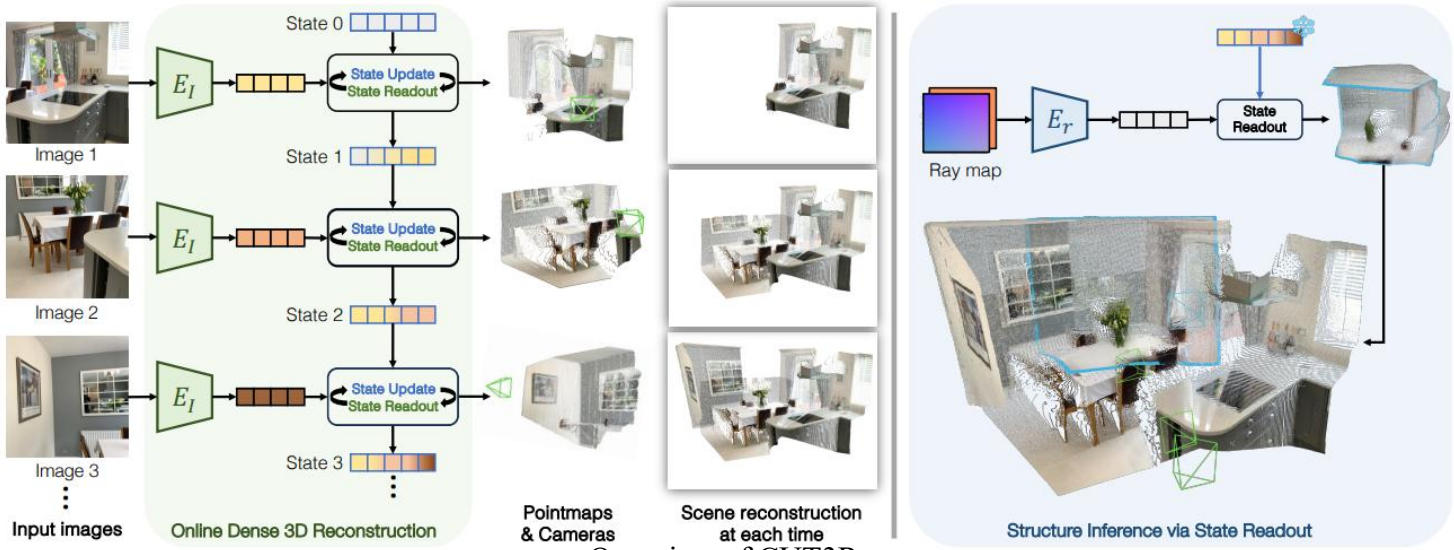
CUT3R

- Problem
  - Lack of Data
    - Point Cloud를 복원하는 기존 SfM이나 MVS 방식은 여러 이미지 간의 픽셀의 매칭을 통해 Point의 3차원 좌표를 복원하나 Image나 Camera 정보가 부족하면 Reconstruction에 어려움을 겪음
  - Optimizing time Consuming
    - Gaussian Splatting 등의 방식은 복원시에 다양한 정보들을 바탕으로 복원과정에서 Optimizing을 진행하여 많은 시간 소모
  - Offline Reconstruction
    - DUSt3R 등의 방식은 빠르고 강력하지만 전체 Image들을 함께 고려하여 복원하므로 Image가 입력되는 즉시 순차적으로 복원하는 실시간(Online) 과정에 적합하지 않음



CUT3R

- Method Overview
  - State Update / Readout
    - 순서대로 입력되는 Image가 State를 반영하여 Camera Parameter와 3D Point Map을 생성함과 동시에 State를 Update
    - 각 3D Point Map이 쌓여가면 전체 3D Scene(Point Cloud)가 복원됨
  - Unseen Structure Inference
    - Update된 State를 통해 가상의 시점에서의 Point Map 복원 가능

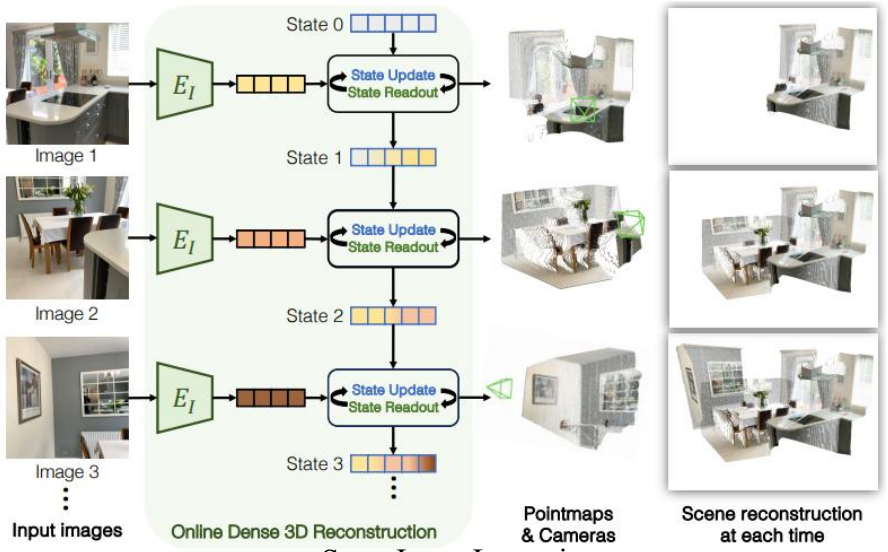


< Overview of CUT3R >

CUT3R

- State-Input Interaction Mechanism
  - State Token
    - 단순한 메모리가 아니라, Inference 과정을 거치면서 누적되는 3D 장면의 기하학적 Context 정보들을 담고 있는 토큰
  - Image Token
    - ViT Encoder를 추출한 현재 Input Image의 토큰 $\boldsymbol{F}_t = \text{Encoder}_i(\boldsymbol{I}_t)$

$\boldsymbol{F}_t$: Current Image Token
$\boldsymbol{I}_t$: Current Input Image



< State-Input Inteaction >

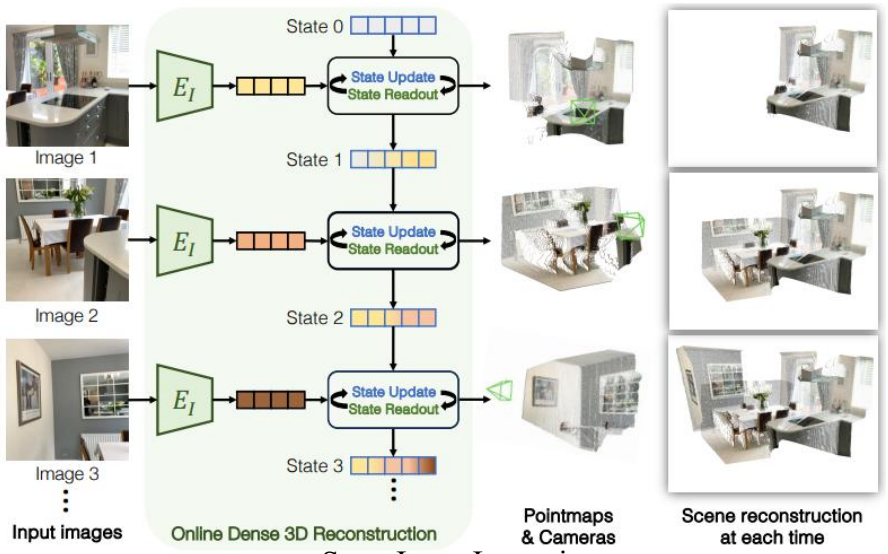
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - State Update/Readout
    - 이전 State에 현재 이미지의 정보를 반영하여 새로운 State를 만드는 State Update
    - 현재 이미지에 이전 State에 저장된 기하학적 Context를 반영하여 기하학적 정보를 담게 되는 State Readout
      - 이 Bidirectional interaction은 State tokens과 Image의 유기적으로 연결하는 CroCo와 DUSt3R의 트랜스포머 디코더를 이용

$$[\boldsymbol{z}'_t, \boldsymbol{F}'_t], \boldsymbol{s}_t = \text{Decoders}\,([\boldsymbol{z}, \boldsymbol{F}_t], \boldsymbol{s}_{t-1})$$

$\boldsymbol{F}$: Image Token
$\boldsymbol{s}$: State Token
$\boldsymbol{z}$: Learnable Pose Token



< State-Input Inteaction >

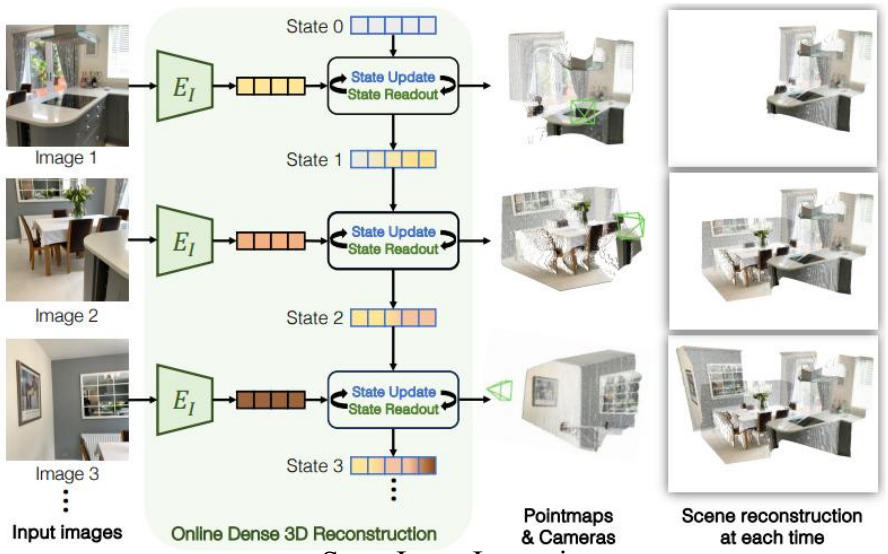
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - State Update/Readout
    - Decoder를 통한 Cross-Attention 이후 State의 기하학적 정보를 반영한 $\boldsymbol{F}'_t$와 $\boldsymbol{z}'_t$ 가 생성
      - Pose Token은 Training과정을 통해 초기값을 학습하고 Decoder에서 Image Token과 State Token의 정보를 반영한 현재의 Pose Token $\boldsymbol{z}'$이 생성

$$[\boldsymbol{z}'_t, \boldsymbol{F}'_t], \boldsymbol{s}_t = \text{Decoders}\,([\boldsymbol{z}, \boldsymbol{F}_t], \boldsymbol{s}_{t-1})$$

$\boldsymbol{F}$: Image Token
$\boldsymbol{s}$: State Token
$\boldsymbol{z}$: Learnable Pose Token



< State-Input Inteaction >

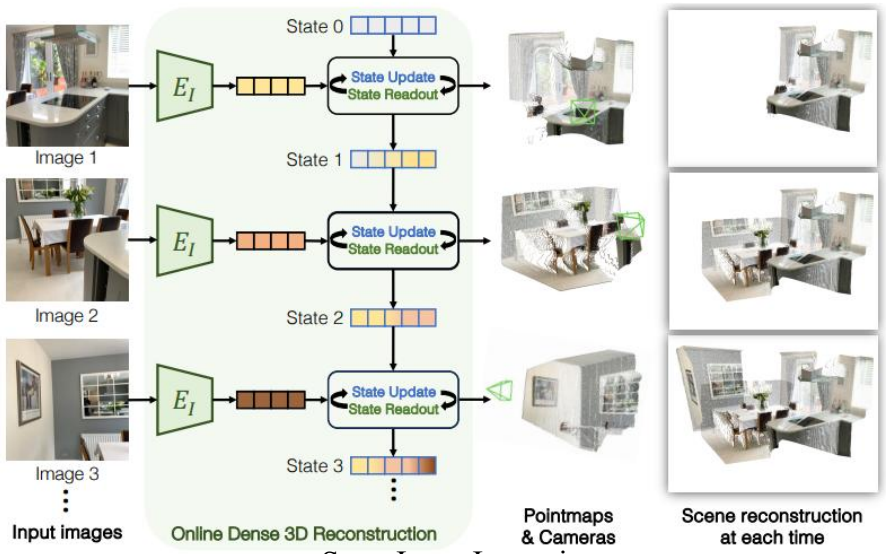
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - Prediction Heads
    - 각 Token으로부터 실제 3D 정보 추출
    - 카메라 기준에서의 point map과 confidence map ($\hat{\boldsymbol{X}}_t^{self}, \boldsymbol{C}_t^{self}$)과 World 좌표계 기준에서의 point map과 confidence map ($\hat{\boldsymbol{X}}_t^{world}, \boldsymbol{C}_t^{world}$)를 추출

$$\hat{\boldsymbol{X}}_t^{self}, \boldsymbol{C}_t^{self} = \text{Head}_{self}(\boldsymbol{F}'_t)$$
$$\hat{\boldsymbol{X}}_t^{world}, \boldsymbol{C}_t^{world} = \text{Head}_{world}(\boldsymbol{F}'_t, \boldsymbol{z}'_t)$$

$\hat{\boldsymbol{X}}$: Predict point map
$\boldsymbol{C}$: Confidence map



< State-Input Inteaction >

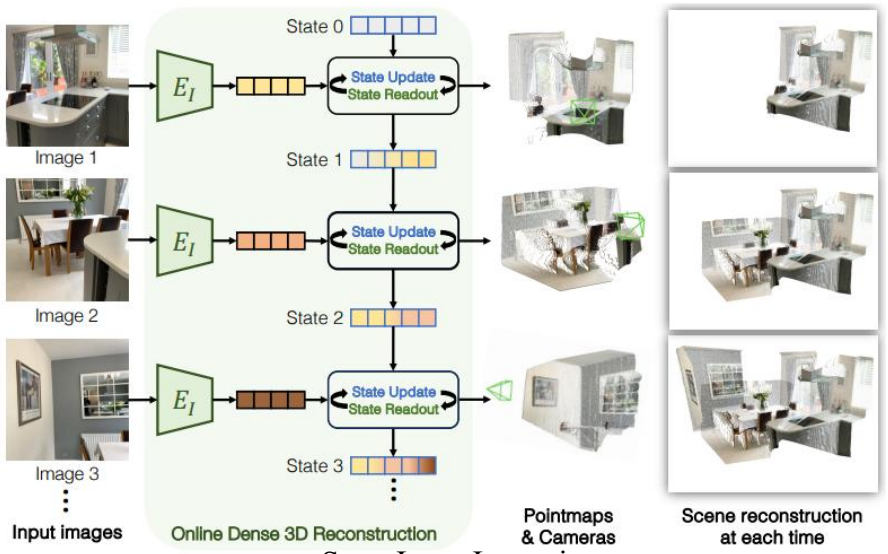
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - Prediction Heads
    - 카메라 좌표계와 World 좌표계 사이의 Relative Transformation인 $\hat{\boldsymbol{P}}_t$를 추가로 추출
      - 간단한 Matrix의 추출이므로 MLP network 이용

$$\hat{\boldsymbol{X}}_t^{self}, \boldsymbol{C}_t^{self} = \text{Head}_{self}(\boldsymbol{F}'_t)$$
$$\hat{\boldsymbol{X}}_t^{world}, \boldsymbol{C}_t^{world} = \text{Head}_{world}(\boldsymbol{F}'_t, \boldsymbol{z}'_t)$$
$$\hat{\boldsymbol{P}}_t = \text{Head}_{pose}(\boldsymbol{z}'_t)$$

$\hat{\boldsymbol{X}}$: Predict point map
$\boldsymbol{C}$: Confidence map
$\hat{\boldsymbol{P}}$: Predict Camera Pose



< State-Input Inteaction >

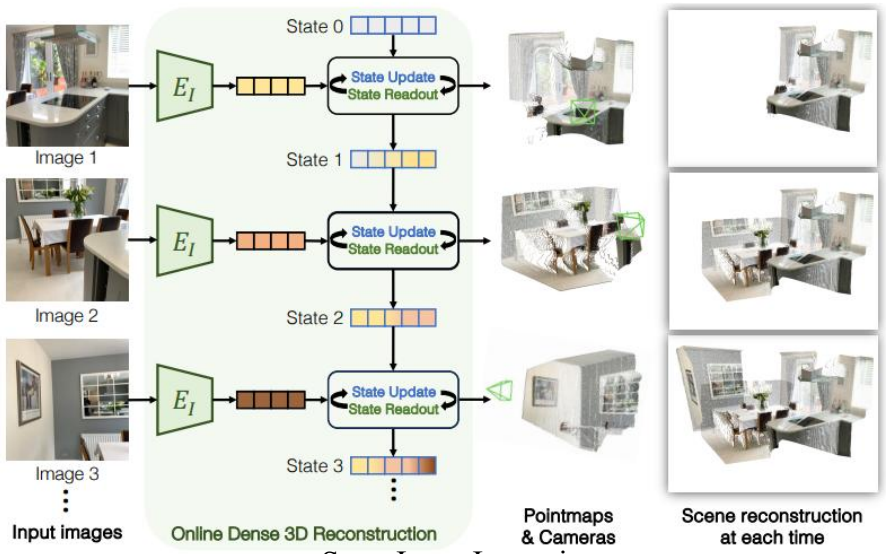
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - Prediction Heads
    - 현재의 카메라 좌표계와 World 좌표계로 변환시킬 수 있는 Transformation인 Camera Pose 정보를 담은 $\hat{\boldsymbol{P}}_t$를 추가로 추출
      - 간단한 Matrix의 추출이므로 MLP network 이용

$$\hat{\boldsymbol{X}}_t^{self}, \boldsymbol{C}_t^{self} = \text{Head}_{self}(\boldsymbol{F}'_t)$$
$$\hat{\boldsymbol{X}}_t^{world}, \boldsymbol{C}_t^{world} = \text{Head}_{world}(\boldsymbol{F}'_t, \boldsymbol{z}'_t)$$
$$\hat{\boldsymbol{P}}_t = \text{Head}_{pose}(\boldsymbol{z}'_t)$$

$\hat{\boldsymbol{X}}$: Predict point map
$\boldsymbol{C}$: Confidence map
$\hat{\boldsymbol{P}}$: Predict Camera Pose

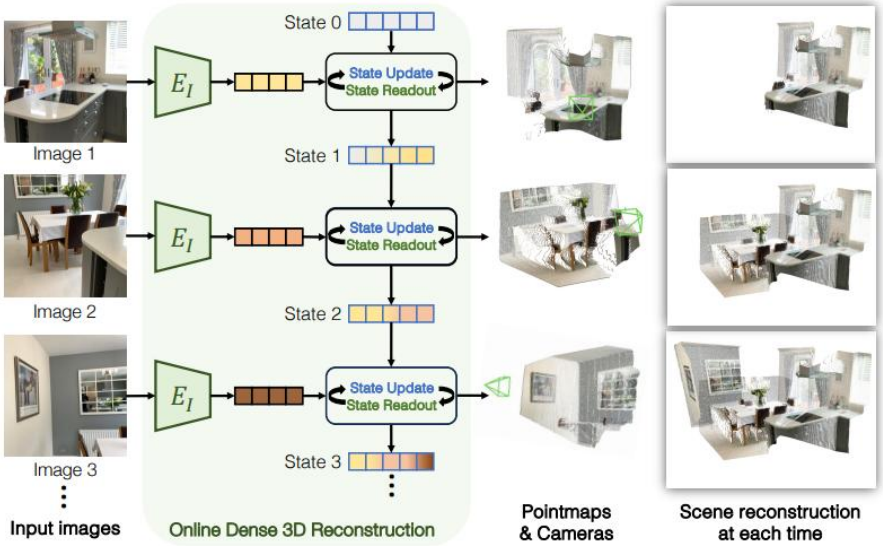


< State-Input Inteaction >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

CUT3R

- State-Input Interaction Mechanism
  - Prediction Heads
    - $\hat{X}_t^{self}$ 와 $\hat{X}_t^{world}$ 그리고 $\hat{P}_t$ 는 서로를 통해 변환가능한, 즉 중복되는 정보를 가진듯 하지만 별개의 Head로 각각 추출
    - 각각의 정보를 추출하므로 학습과정서 Depth map 또는 Camera Parameter가 부족할 때 일부 Annotation만으로 학습이 가능
      - 다양한 데이터셋을 활용 가능



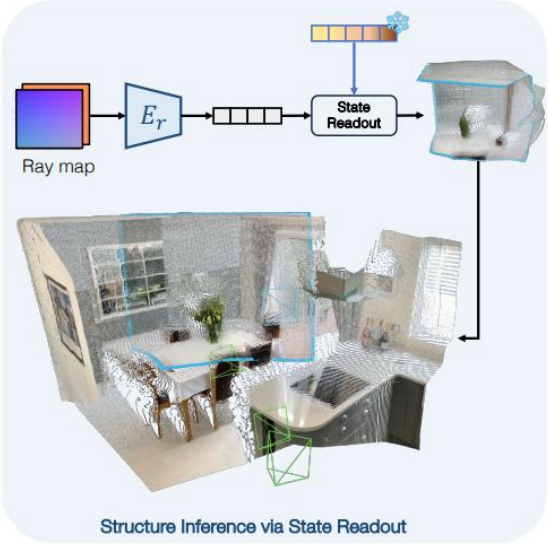
< State-Input Inteaction >

$$\hat{X}_t^{self}, C_t^{self} = \text{Head}_{self}(F'_t)$$
$$\hat{X}_t^{world}, C_t^{world} = \text{Head}_{world}(F'_t, z'_t)$$
$$\hat{P}_t = \text{Head}_{pose}(z'_t)$$

$\hat{X}$: Predict point map
C: Confidence map
$\hat{P}$: Predict Camera Pose

CUT3R

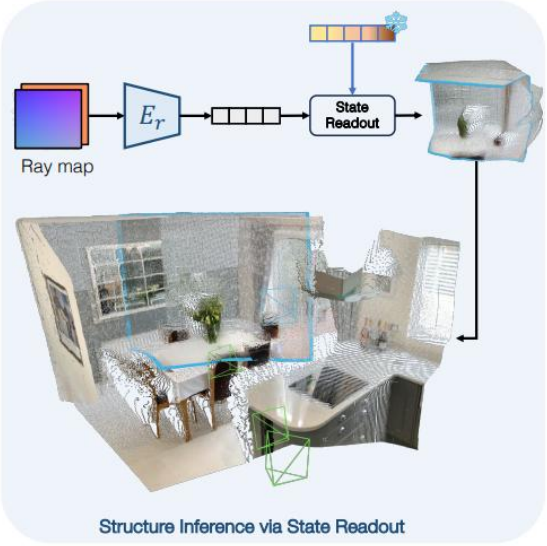
- Querying the State with Unseen Views
  - Unseen Structure Inference
    - 가상의 카메라의 파라미터 정보를 담고 있는 6channel의 이미지 Raymap을 통해 가상의 시점에서의 Point Map 복원
      - Raymap은 카메라의 origin좌표와 각 픽셀에서 카메라로의 빛(ray)의 방향을 6channel로 나타냄



< Unseen Structure Inference >

CUT3R

- Querying the State with Unseen Views
  - Unseen Structure Inference
    - Query raymap으로 부터 인코더를 통해 Raymap Token $\boldsymbol{F}_r$ 을 생성
    - 앞선 과정에서 생성된 State와 State Readout 과정을 통해 Unseen View에서의 Point Map 복원
      - State가 3D Context를 잘 담고 있다는 증거
      - Color 정보는 Color Head를 이용



< Unseen Structure Inference >

$$\hat{\boldsymbol{I}}_r = \text{Head}_{color}(\boldsymbol{F}'_r)$$

$\boldsymbol{F}_r$: Raymap Token
$\hat{\boldsymbol{I}}_r$: Predict Image(Color)

CUT3R

- Training
  - Input Sequence
    - N개의 이미지 시퀀스를 연속된 비디오, 순서 없는 사진 모음(dynamic 제외) 등 다양한 Sequence를 학습
    - 일정 확률로 해당시점의 Raymap을 Input으로 이용하여 3D Context 학습 강화
  - Training Sequence
    - 1st stage: Static Scene 4 views 이미지로 기본적인 Prior 학습
    - 2nd stage: Dynamic Scene과 Partial annotations 데이터셋도 활용하여 일반화 성능을 높임
    - 3rd stage: 최대 512pixel로 해상도를 높여 정밀도를 높임
    - 4th stage: 최대 64 views의 Long Context를 학습해 State의 일관성을 유지한다
      - 전에 학습된 Encoder는 영향을 주지 않으므로 Freeze

CUT3R

- Experiments
  - Quantitative Result
    - 빠른 속도로 비교적 정밀한 Depth와 카메라 포즈가 추정되었다

Alignment	Method	Optim.	Onl.	Sintel Abs Rel ↓	Sintel δ<1.25 ↑	BONN Abs Rel ↓	BONN δ<1.25 ↑	KITTI Abs Rel ↓	KITTI δ <1.25 ↑	FPS
Per-sequence scale	DUSt3R-GA [107]	✓		0.656	45.2	0.155	83.3	0.144	81.3	0.76
	MASt3R-GA [51]	✓		0.641	43.9	0.252	70.1	0.183	74.5	0.31
	MonST3R-GA [125]	✓		**0.378**	**55.8**	**0.067**	**96.3**	0.168	74.4	0.35
	Spann3R [101]		✓	0.622	42.6	0.144	81.3	0.198	73.7	13.55
	Ours		✓	0.421	47.9	0.078	93.7	**0.118**	**88.1**	16.58
Metric scale	MASt3R-GA [51]	✓		**1.022**	14.3	0.272	70.6	0.467	15.2	0.31
	Ours		✓	1.029	**23.8**	**0.103**	**88.5**	**0.122**	**85.5**	16.58

Table 2. **Video Depth Evaluation**. We report scale-invariant depth and metric depth accuracy on Sintel, Bonn, and KITTI datasets. Methods requiring global alignment are marked "GA", while "Optim." and "Onl." indicate optimization-based and online methods, respectively. We also report the FPS on KITTI dataset using 512× 144 image resolution for all methods on an A100 GPU, except Spann3R which only supports 224×224 inputs. We present a subset of baselines here; please refer to the supplementary material for full comparisons.

Method	Optim.	Onl.	Sintel ATE ↓	Sintel RPE trans ↓	Sintel RPE rot ↓	TUM-dynamics ATE ↓	TUM-dynamics RPE trans ↓	TUM-dynamics RPE rot ↓	ScanNet ATE ↓	ScanNet RPE trans ↓	ScanNet RPE rot ↓
Particle-SfM [129]	✓		0.129	0.031	**0.535**	-	-	-	0.136	0.023	0.836
Robust-CVD [47]	✓		0.360	0.154	3.443	0.153	0.026	3.528	0.227	0.064	7.374
CasualSAM [128]	✓		0.141	**0.035**	0.615	0.071	**0.010**	1.712	0.158	0.034	1.618
DUSt3R-GA [107]	✓		0.417	0.250	5.796	0.083	0.017	3.567	0.081	0.028	0.784
MASt3R-GA [51]	✓		0.185	0.060	1.496	**0.038**	0.012	**0.448**	0.078	0.020	**0.475**
MonST3R-GA [125]	✓		**0.111**	0.044	0.869	0.098	0.019	0.935	**0.077**	**0.018**	0.529
DUSt3R [107]		✓	0.290	0.132	7.869	0.140	0.106	3.286	0.246	0.108	8.210
Spann3R [101]		✓	0.329	0.110	4.471	0.056	0.021	0.591	**0.096**	0.023	0.661
Ours		✓	**0.213**	**0.066**	**0.621**	**0.046**	**0.015**	**0.473**	0.099	**0.022**	**0.600**

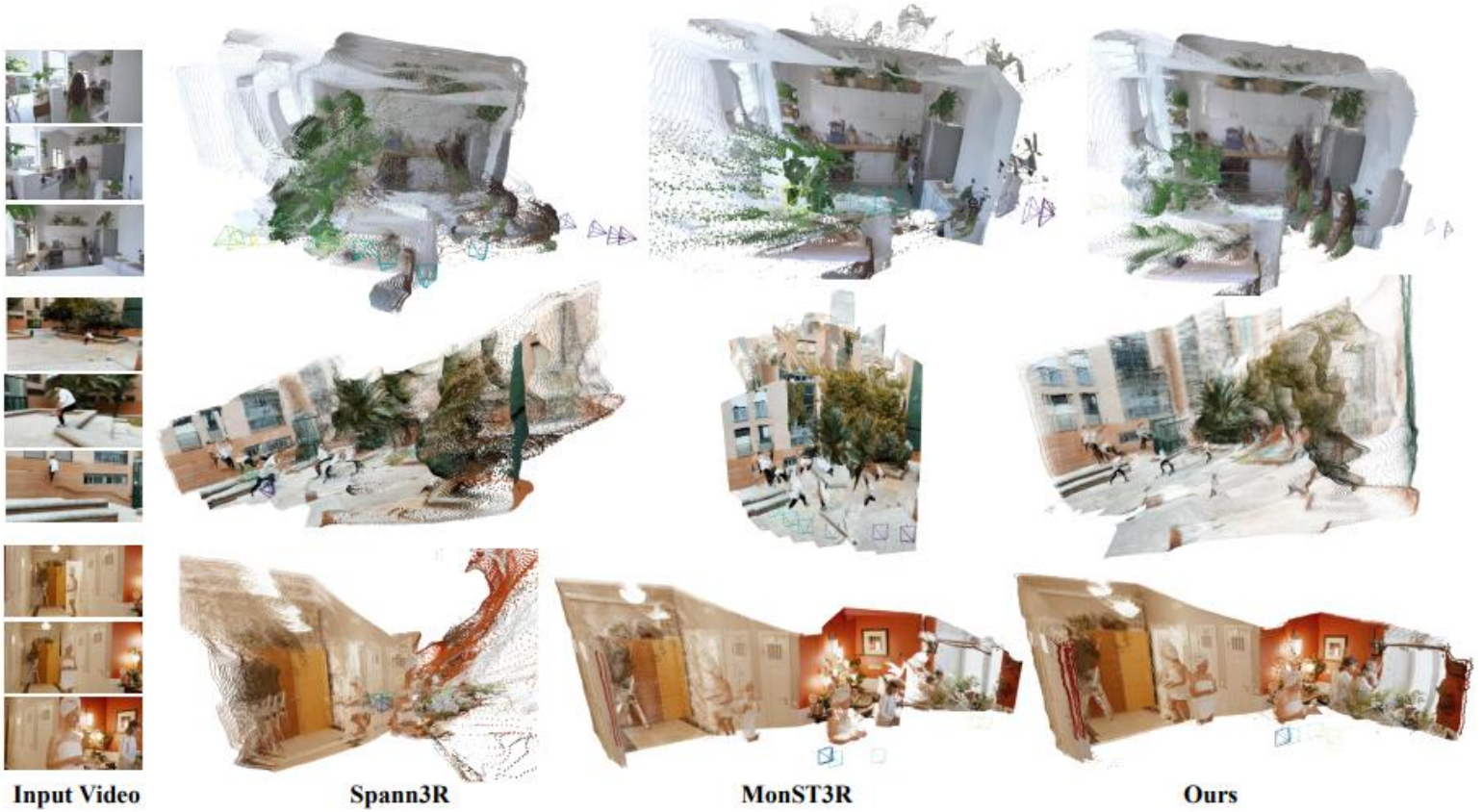
Table 3. **Evaluation on Camera Pose Estimation** on Sintel [12], TUM-dynamic [89], and ScanNet [19] datasets. Our method achieves the best overall performance among all online methods.

Abs Rel: Detph 상대 오차율
$\delta < 1.25$: Depth비 1.25 이하 비율

ATE: Align 후 카메라 경로 평균오차
RPE: 프레임 간의 카메라 이동/회전 오차

CUT3R

- Experiments
  - Qualitive Result



CUT3R

- Experiments
  - Qualitive Result
    - Unseen Structure Reconstruction





Human3R: Everyone Everywhere All at Once [ICLR 26]

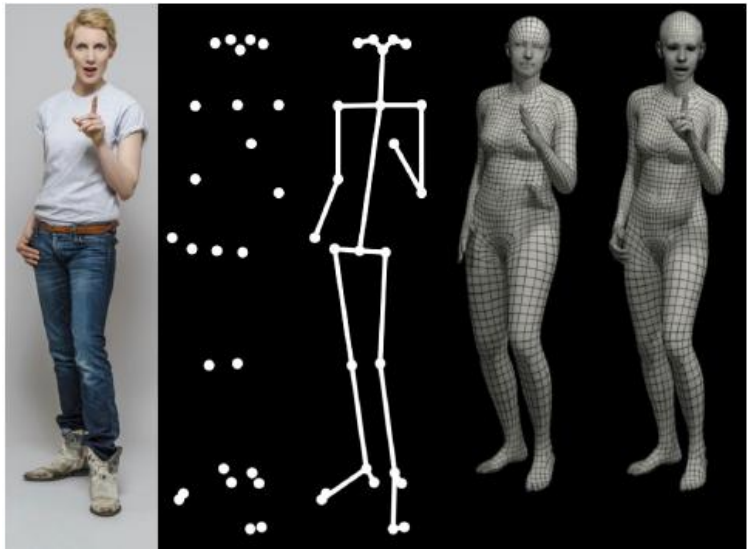
Human3R

- Human3R
  - 영상에서 3D Point Cloud와 Human Mesh를 동시에 복원하는 모델



Human3R

- Background
  - SMPL-X
    - 간단한 파라미터의 조합으로 사람의 전체 포즈, 체형, 얼굴 등을 표현하는 모델
      - 키, 몸무게, 체격 등의 특징을 가지는 β와 각 관절의 각도를 나타내는 θ, 얼굴 표정을 나타내는 α로 표현



< SMPL-X >

Human3R

- Problem
  - Multi-stage/model/shot
    - Human과 Scene을 개별 모델로 복원하고 Contact 등 여러 조건에 맞게 Optimizing으로 Align을 맞춤
      - 복잡한 Pipeline, 많은 연산/시간 소요, 별도의 Human Detection/Crop 필요
  - Heavy Dependencies
    - Metric Depth Estimators, 3D Reconstruction Models, Camera Pose and Intrinsic 등 다양한 모델의 Pre-processing의 의존도가 높음
      - Non real time inference, 오차 누적

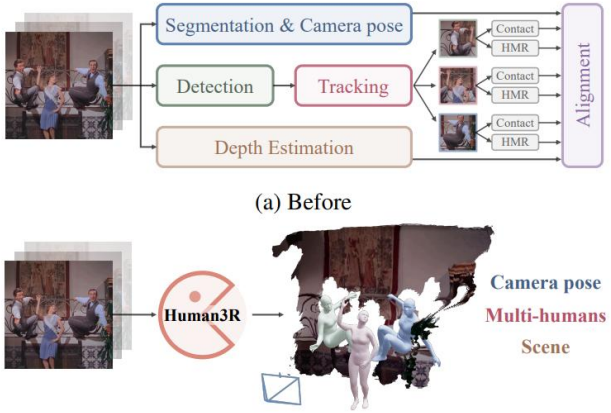


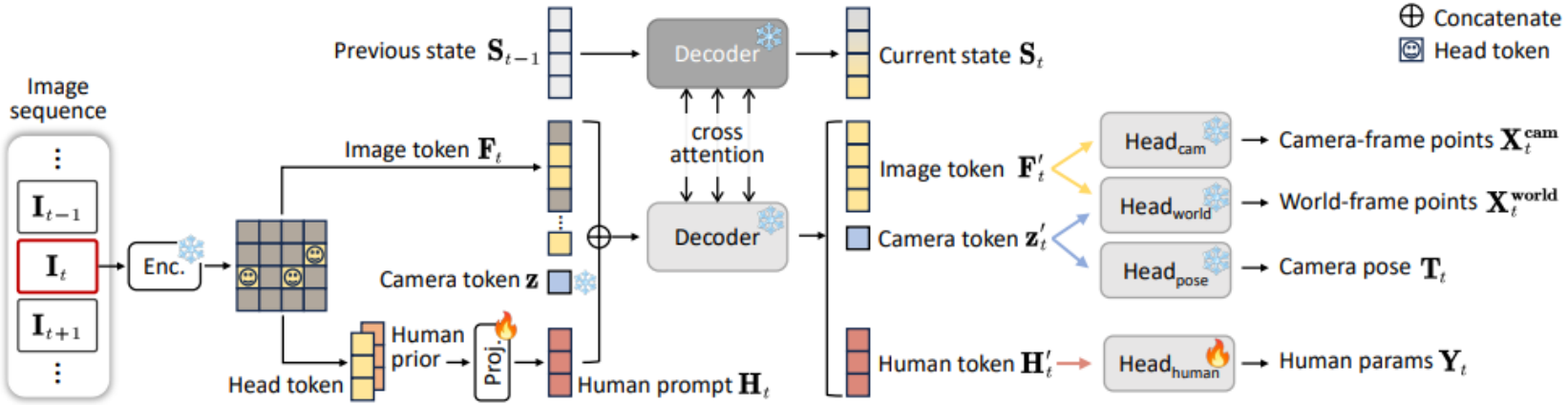
Figure 3: Multi-stage vs. One-stage.

Human3R

- Goal (All at Once)
  - One Model
    - 별도의 모델 없이 Humans, scene, camera를 하나의 모델로
  - One Stage
    - Feed-forward 방식으로 실시간으로
  - One Shot
    - Multiple people을 single forward pass로
  - One GPU, One Day
    - 효율적으로 하나의 GPU로 하루만에 학습

Human3R

- Method
  - CUT3R를 베이스로 Human Token을 추가
    - Human-related layer만 Fine-tuning



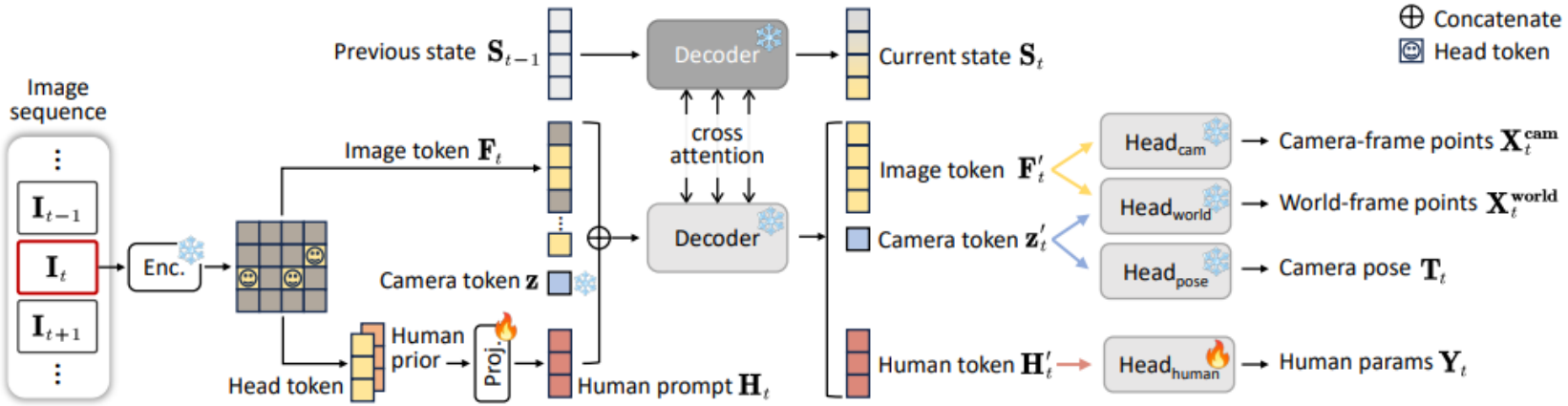
< Architecture Overview >

Human3R

- Method
  - Global Human-Scene Reconstruction
    - Encoder를 통과한 Image Token의 각 Patch에서 Head를 검출
    - 이 작은 Patch를 Head Token으로 삼아 이를 Human Prompt로 Projection

$$\boldsymbol{H}_t = \text{Head}_{\text{projection}}(\boldsymbol{F}_t^{\boldsymbol{u}})$$

$\boldsymbol{H}_t$: Human prompt
$\boldsymbol{F}_t$: Image token
u: Head가 Detecting된 PATCH



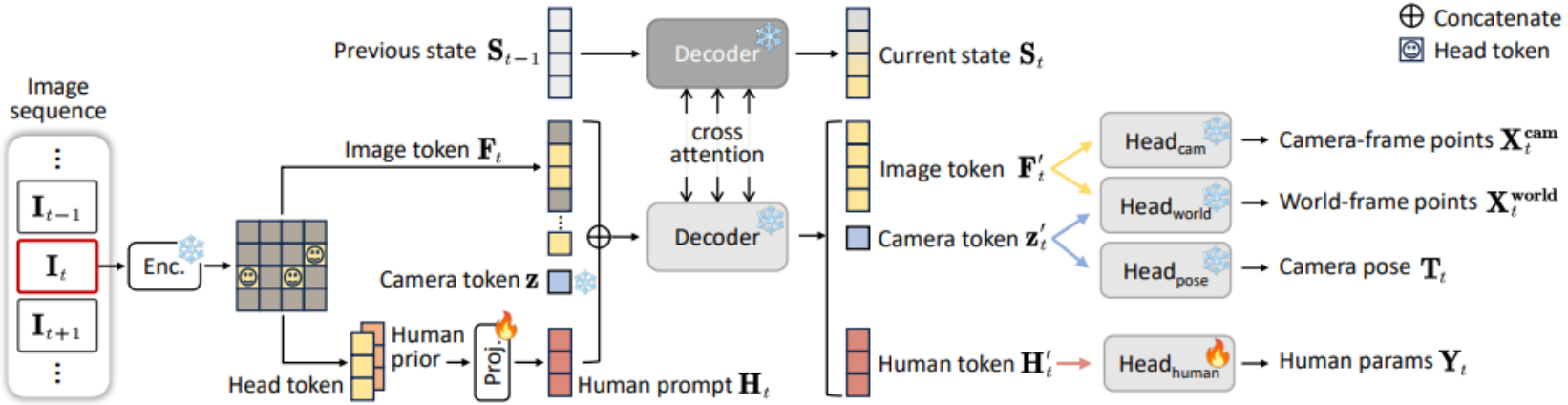
< Architecture Overview >

Human3R

- Method
  - Global Human-Scene Reconstruction
    - Human prompt는 Decoder를 Image/Camera Token과 함께 통과하면 Image, Camera, State의 정보를 통해 Local Pose 및 Global Origin/Rotation 정보를 가지는 Human Token 생성
    - Human Token은 Human Head를 통해 SMPL-X 파라미터와 위치를 복원

$$[\boldsymbol{F}'_t, \boldsymbol{z}'_t, \boldsymbol{H}'_t], \boldsymbol{s}_t = \text{Decoders}\left([\boldsymbol{F}_t, \boldsymbol{z}, \boldsymbol{H}_t], \boldsymbol{s}_{t-1}\right)$$
$$\boldsymbol{Y}_t = \text{Head}_{\text{human}}(\boldsymbol{H}'_t)$$

$\boldsymbol{H}_t$: Human prompt
$\boldsymbol{H}'_t$: Human token
$\boldsymbol{F}_t$: Image token
$\boldsymbol{z}$: Pose token
$\boldsymbol{Y}_t$:Human Parameters
$\boldsymbol{Y}_t = \{(\theta, \beta, \alpha, P^{cam})_t\}_n$



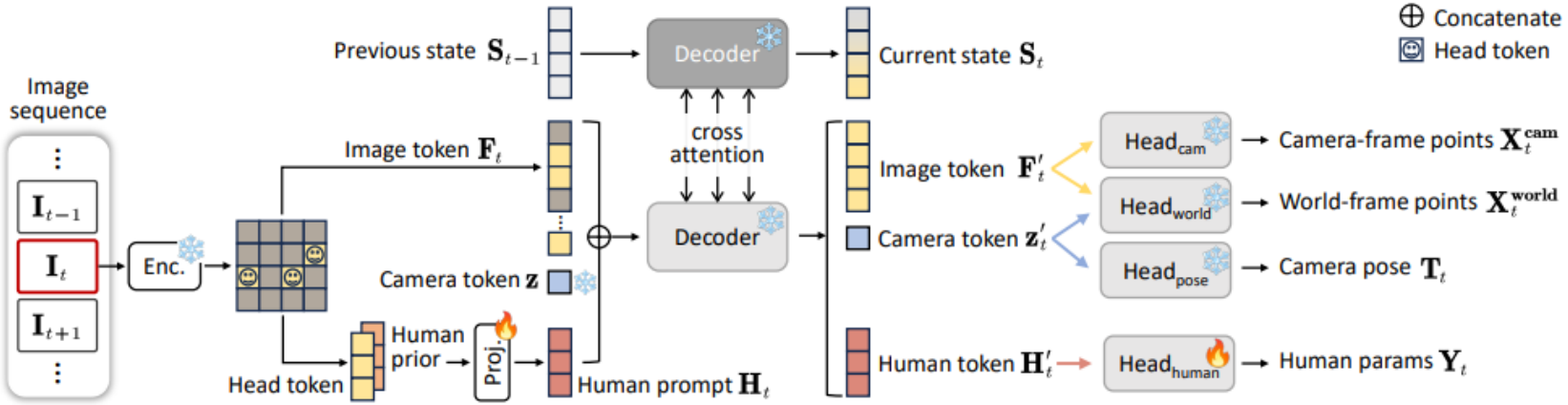
< Architecture Overview >

Human3R

- Method
  - Human Prior
    - CUT3R는 많은 양의 데이터 셋을 학습하였지만 human에 관한 정보는 부족
    - Head Token $\boldsymbol{H}^u$에 Human Prior관련 토큰을 추가
    - Multi-HMR이라는 One Shot으로 Human Pose Reconstruction을 하는 대량의 Human Prior가 학습된 Multi-HMR 모델의 Encoder를 통해 Human Token을 추가 생성

$$\boldsymbol{F}_{HMR} = \text{Encoder}_{HMR}(\boldsymbol{I})$$
$$\boldsymbol{Y}_t = \text{Head}_{projection}(\boldsymbol{H}^u \oplus \boldsymbol{F}^u_{HMR})$$

⊕: Concatenation
$\boldsymbol{I}$: Input Image
$\boldsymbol{F}_t$: Image token
$\boldsymbol{Y}_t$:Human Parameters

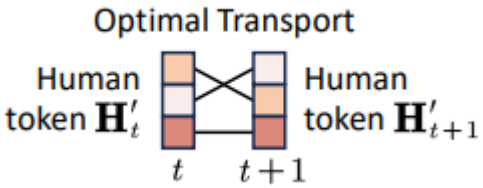


< Architecture Overview >

Human3R

- Method
  - Human Tracking
    - 각 Human Prompt, Token에는 각각의 사람을 인식할 수 있을 정도의 정보가 보존
    - 이전 Human Token 정보를 저장하였다가 새 Input에서의 Human Token과 비교하여 토큰 간의 L2 Distance가 Threshold이하인 가장 작은 Human Token끼리 매칭



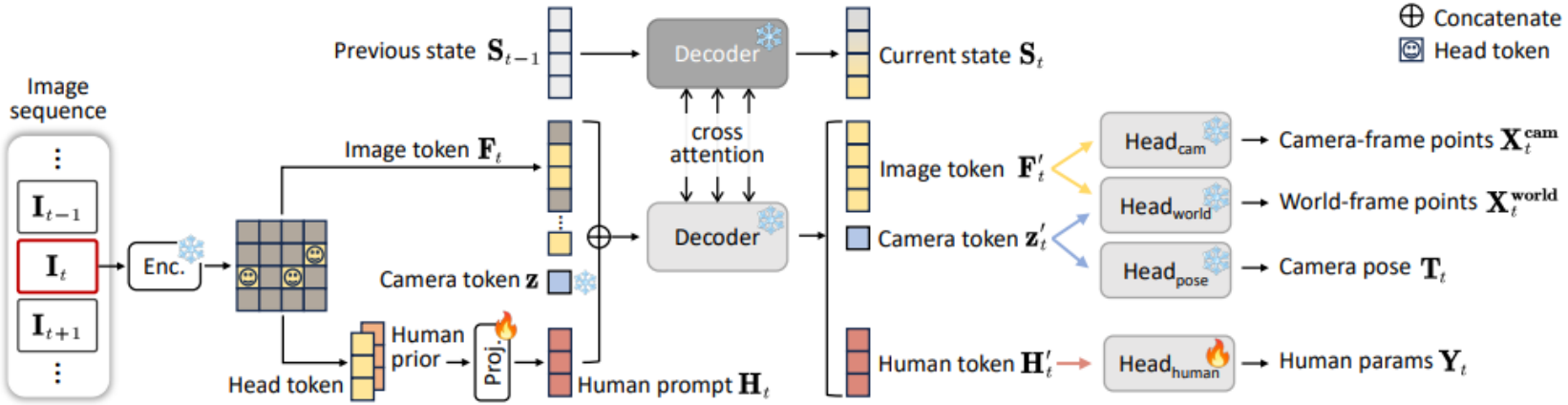


< Human Tracking >



Human3R

- Training
  - Human Prompt Tuning
    - 메인 백본을 그대로 유지한 채, Prior(Multi-HMR) 등을 이용해 일부 토큰을 추가해 Prompt를 만들어 추가한 Layer만을 학습하여 매우 빠르고 간단히 학습이 가능
      - 48GB GPU로 하루만에 학습



< Architecture Overview >

Human3R

- Experiments
  - Quantitative result
    - 8GB GPU로 15FPS 추론 가능
    - 다른 모델보다 적은 조건에도 불구하고 우수한 Human Local Pose / Global Pose를 추정하였다

Category	Method	Crop-free	Detection-free	Intrinsic-free	3DPW (14) PA-MPJPE ↓	3DPW (14) MPJPE ↓	3DPW (14) PVE ↓	EMDB-1 (24) PA-MPJPE ↓	EMDB-1 (24) MPJPE ↓	EMDB-1 (24) PVE ↓
Multi-stage	CLIFF [52]	✗	✗	✗	43.0	69.0	81.2	68.3	103.3	123.7
	HMR2.0a [31]	✗	✗	✓	44.4	69.8	82.2	61.5	97.8	120.0
	TokenHMR [25]	✗	✗	✓	44.3	71.0	84.6	55.6	91.7	109.4
	CameraHMR [62]	✗	✗	✓	38.5	62.1	72.9	43.7	73.0	85.4
	NLF [76]	✗	✗	✗	37.3	60.3	71.4	41.2	**69.6**	**82.4**
	PromptHMR [100]	✓	✗	✗	**36.6**	**58.7**	**69.4**	**41.0**	71.7	84.5
One-stage	BEV [88]	✓	✓	✓	46.9	78.5	92.3	70.9	112.2	133.4
	Multi-HMR [3]	✓	✓	✗	45.9	73.1	87.1	50.1	81.6	95.7
	Human3R	✓	✓	✓	**44.1**	**71.2**	**84.9**	**48.5**	**73.9**	**86.0**

Table 1: **Evaluation of local human mesh reconstruction** on 3DPW [94] and EMDB-1 [39] datasets.

Category	Method	Preprocessed Input (✓ = not required) Detection	Tracking	LocalHuman	Camera	Mask	Depth	Contact	Output GlobalHuman	CameraPose	Scene	EMDB-2 (24) WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	RICH (24) WA-MPJPE ↓	W-MPJPE ↓	RTE ↓
Offline	GLAMR [110]	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	280.8	726.6	11.4	129.4	236.2	3.8
	SLAHMR [109]	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	326.9	776.1	10.2	132.2	237.1	6.4
	COIN [48]	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	152.8	407.3	3.5	169.5	254.5	-
	GVHMR [80]	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	111.0	276.5	2.0	**78.8**	**126.3**	**2.4**
	TRAM [99]	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	76.4	222.4	1.4	127.8	238.0	6.0
	JOSH [53]	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	**68.9**	**174.7**	**1.3**	89.0	132.5	3.0
Online	TRACE [89]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	529.0	1702.3	17.7	238.1	925.4	610.4
	WHAM [82]	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	135.6	354.8	6.0	**108.4**	196.1	4.5
	JOSH3R [53]	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	220.0	661.7	13.1	-	-	-
	Human3R	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**112.2**	**267.9**	**2.2**	110.0	**184.9**	**3.3**

Table 2: **Evaluation of global human motion estimation** on EMDB-2 [39] and RICH [33] datasets.

MPJPE: Root Joint 기준Joint 평균 오차
PVE: Mesh 전체 오차
W-MPJPE: World 좌표계 기준 MPJPE
RTE: Human Position 오차

Human3R

- Experiments
  - Qualitative result



