From Depth Estimation to 4D Scene Understanding

RILS LAB

Robot Intelligence Learning & System

김채연

2026. 07. 16

Paper List

1. **GeoDepth**
   Self-Supervised Monocular Depth Estimation

2. **EgoMono4D**
   Self-Supervised Monocular 4D Scene Reconstruction

Depth Estimation → 4D Reconstruction

01

GeoDepth: From Point-to-Depth to Plane-to-Depth Modeling for Self-Supervised Monocular Depth Estimation

Haifeng Wu, Shuhang Gu, Lixin Duan, Wen Li

CVPR / 2025

GeoDepth

Abstract
기존 Self-Supervised MDE은 각 pixel을 독립적인 point로 처리하고 pixel별 depth를 직접 예측

Problem
- **Depth Discontinuity:** 동일 평면에서도 깊이값이 갑자기 변함
- **Low-texture Region:** 벽·도로처럼 무늬가 적은 영역에서 오차 발생

Contribution
새로운 **Plane-to-Depth Modeling** 제안

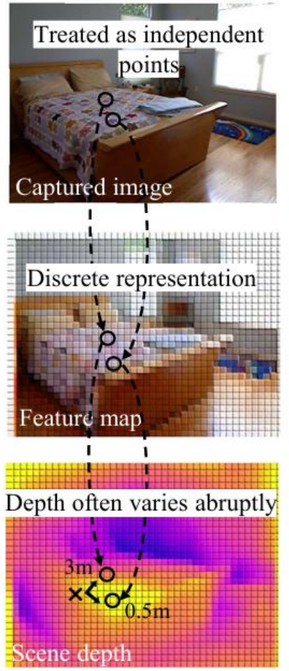
Monocular Depth Estimation

Depth Estimation
한 장의 RGB 이미지에서 각 픽셀까지의 거리를 예측하여 **2D Image를 3D Scene Structure로 변환**

Applications
- Autonomous Driving: 차량·보행자와의 거리 판단
- Robotics: 장애물 회피와 물체 조작
- 3D Reconstruction: 공간 구조 복원
- AR: 실제 공간에 가상 물체 배치

Why Self-supervised?
Dense Depth Label은 LiDAR 등의 센서가 필요해 수집 비용이 큼
→ 정답 깊이 없이 **Video 또는 Stereo Image**로 학습

MDE
- 낮은 센서 비용: **RGB Camera** 한 대만으로 깊이 추정
- 간단한 전처리: **Stereo** 방식의 두 카메라 사이 거리 · 측정 각도 보정 과정이 적음
- 높은 활용성: 경량 모델 사용 시 제한된 메모리와 실시간 환경에도 적용 가능



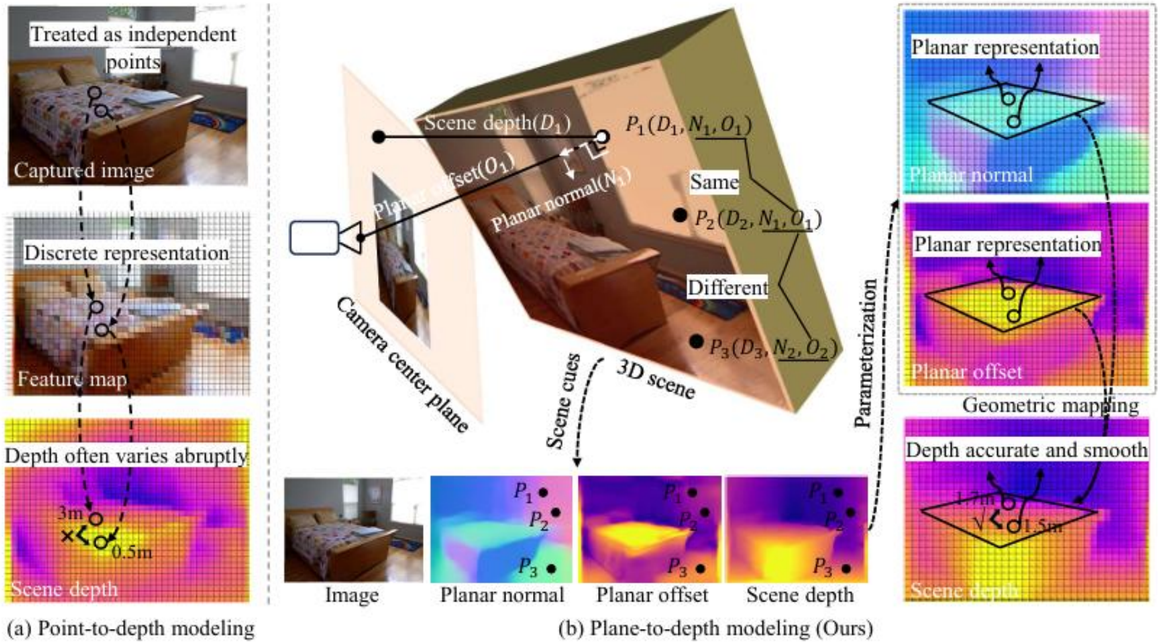
(a) Point-to-depth modeling

From Point-to-Depth to Plane-to-Depth

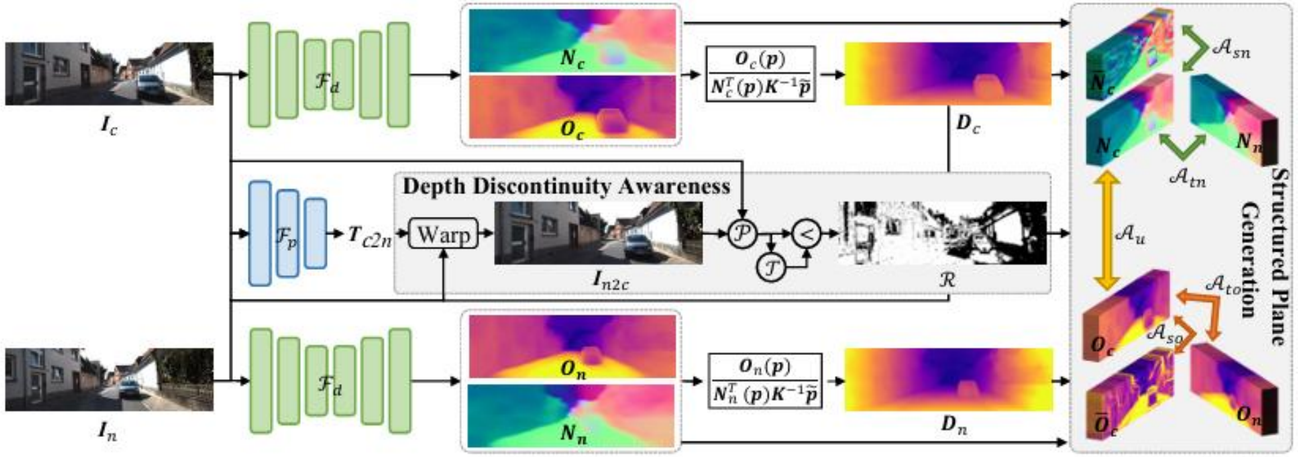
Point-to-Depth
각 픽셀을 독립적인 점으로 처리하여 깊이를 직접 예측
- 같은 벽이나 바닥에서도 깊이값이 갑자기 달라질 수 있음
- 텍스처가 부족한 영역에서 Depth Prediction이 불안정

Plane-to-Depth
3D 장면을 다양한 크기의 **Plane 집합**으로 모델링
- **Planar Normal:** 평면이 향하는 방향
- **Planar Offset:** 카메라 중심에서 평면까지의 거리
- 같은 평면의 픽셀은 동일한 **Normal과 Offset**을 공유

GeoDepth
Planar Normal + Planar Offset → Scene Depth



GeoDepth Overview



1. Plane Prediction
현재 프레임과 인접 프레임에서 **Planar Normal** N과 **Planar Offset** O을 예측

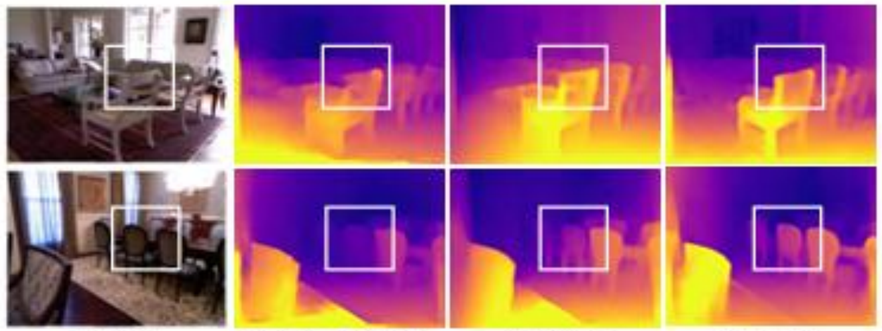
2. Plane-to-Depth
예측한 평면의 방향 N과 거리 O를 이용해 , 각 픽셀의 카메라 ray가 평면과 만나는 지점을 계산하고 Depth Map 생성

3. Depth Discontinuity Awareness
인접 프레임을 현재 시점으로 변환해 실제 이미지와 비교하고, 차이가 큰 영역을 **Depth** 불연속 후보로 탐지

4. Structured Plane Generation
같은 평면의 **Normal과 Offset이 일관되도록** 학습하여 평면 영역의 **Depth**가 자연스럽게 이어지도록 함

Experimental Results & Takeaway

Method	P2D	SPG	DDA	RMSE ↓	Sq Rel↓	$\delta<1.25$↑	#Params
Baseline				4.471	0.751	0.895	9.98M
+P2D	✓			4.436	0.740	0.896	10.0M
+P2D+SPG	✓	✓		4.412	0.722	0.896	10.0M
GeoDepth	✓	✓	✓	**4.381**	**0.694**	**0.897**	10.0M



(a) Input (b) StructDepth (c) PLNet (d) Ours

Ablation Study

Baseline: RMSE **4.471**

- Plane-to-Depth: **4.436**
- Structured Plane Generation: **4.412**
- Depth Discontinuity Awareness: **4.381**

파라미터 수는 약 **10M**으로 거의 증가하지 않으면서 성능 향상

Qualitative Results

- 가구에서 평면 영역이 더 연속적
- 물체 경계가 더 선명하게 보존

Takeaway

GeoDepth는 pixel-wise depth 예측의 한계를 plane structure로 보완하여, 평면 영역에서 더 안정적이고 연속적인 depth map을 생성

02

Self-Supervised Monocular 4D Scene Reconstruction for Egocentric Videos

Chengbo Yuan, Geng Chen, Li Yi, Yang Gao

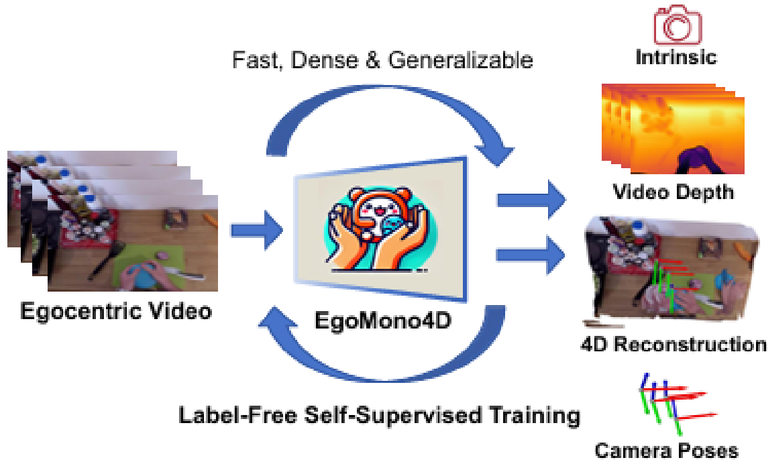
ICCV / 2025

EgoMono4D

Abstract
1인칭 영상은 손과 물체의 상호작용 정보를 풍부하게 포함하지만, 카메라와 물체가 동시에 움직이기 때문에 4D scene reconstruction이 어려움

Problem
- 카메라, 손, 물체가 동시에 움직여 프레임 간 **3D 구조 정렬이 어려움**
- **Gaussian Splatting 기반 방법**은 각 영상마다 별도 최적화가 필요해 속도가 느림
- Supervised 4D reconstruction은 **고품질 depth·pose label**이 필요

목표
정답 라벨 없이 빠르고 조밀한 4D scene reconstruction



From Depth to 4D Reconstruction

Monocular Depth Estimation
한 장의 이미지에서 픽셀별 거리 추정

RGB *Image* → *Depth Map*

4D Scene Reconstruction
영상의 모든 프레임을 하나의 전역 좌표계에 정렬하여 시간에 따른 **3D 구조와 움직임**을 함께 복원

Video → *3D Pointcloud Sequence*

Connection with GeoDepth
- GeoDepth: 한 프레임의 정확하고 연속적인 Depth 추정
- EgoMono4D: 여러 프레임의 Depth와 Camera Motion을 결합해 4D 장면 복원

What Should Be Estimated?

What Should Be Estimated?
입력은 T개의 RGB 프레임으로 구성된 단안 1인칭 영상

$$\{I_0, I_1, \ldots, I_{T-1}\}$$

각 프레임의 모든 픽셀을 전역 3D 좌표로 변환하기 위해 다음 정보를 추정

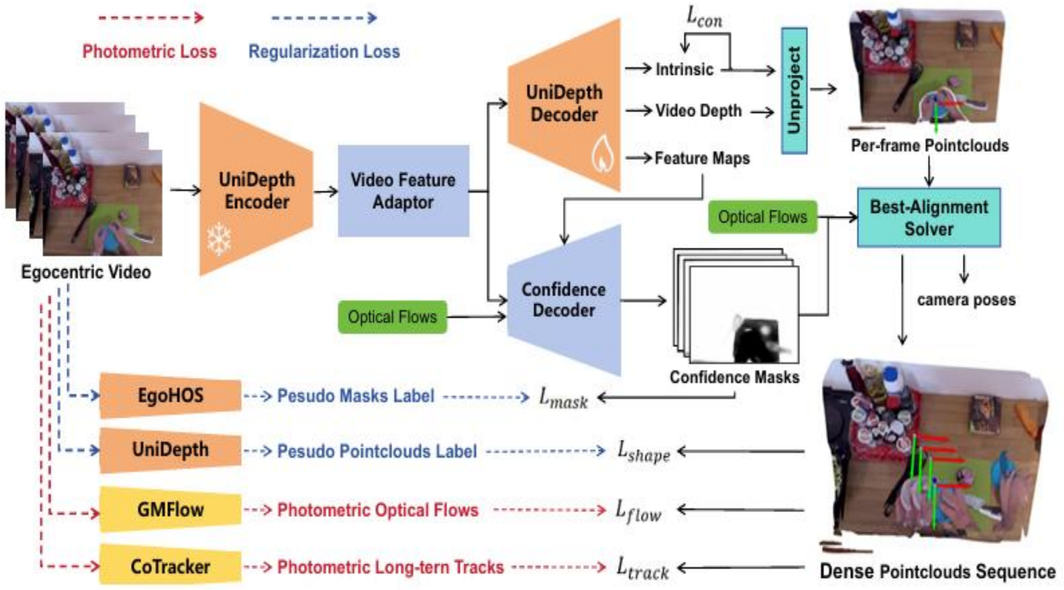
1. Video Depth
각 픽셀까지의 거리
2. Camera Intrinsic
초점거리와 영상 중심 등 카메라 내부 정보
3. Camera Pose
각 프레임에서 카메라의 위치와 방향

Final Output

$$\text{Depth} + \textit{Intrinsic} + \textit{Pose} \rightarrow \textit{Dense Pointcloud Sequence}$$

모든 프레임의 3D 점을 하나의 전역 좌표계에 정렬

EgoMono4D Overview



1. **Video Feature Extraction**
   각 프레임에서 이미지 특징 추출
2. **Multi-frame Feature Fusion**
   프레임 사이의 시간적 정보를 결합
3. **Joint Prediction**
   video depth, camera intrinsic, confidence mask 예측
4. **Camera Pose Estimation**
   프레임별 pointcloud를 정렬해 camera pose 계산
5. **4D Reconstruction**
   모든 프레임을 전역 좌표계에 배치해 dense pointcloud sequence 생성

Experimental Results & Conclusion

	HOI4D				H2O				POV-Surgery†				ARCTIC-HOI†			
	CD ↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑	CD ↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑	CD ↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑	CD ↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑
Modularized Version	8.9	15.4	38.2	68.0	**4.7**	47.8	81.5	94.2	223.3	3.6	9.6	19.1	5.8	12.6	33.8	63.2
DS+UniDepth [66]	6.7	23.2	53.4	79.9	5.1	36.5	75.0	93.2	39.1	7.7	20.0	38.7	2.9	22.2	**60.2**	84.7
DUSt3R [75]	8.6	24.0	53.2	76.8	8.8	23.1	56.0	82.7	55.4	9.7	24.7	45.1	3.2	20.8	53.7	83.1
MonSt3R [87]	7.6	21.4	50.8	77.9	14.7	15.1	42.4	70.0	94.5	6.5	17.8	34.8	5.4	9.9	31.8	65.0
Align3R [38]	7.1	22.3	53.4	78.9	11.5	20.0	45.4	72.0	41.1	7.9	21.0	40.4	4.5	14.2	41.5	75.8
CUT3R [73]	7.5	20.1	47.5	74.3	7.4	17.2	57.2	93.1	107.6	4.7	12.8	25.6	4.5	18.7	48.2	77.9
EgoMono4D (Ours)	**5.9**	**27.9**	**59.6**	**83.1**	5.1	**54.2**	**83.9**	**94.4**	**33.8**	**13.5**	**32.0**	**53.9**	**2.8**	**24.1**	57.5	**86.2**

Results

- HOI4D, H2O에서 기존 방법보다 우수한 dense pointcloud reconstruction 성능
- POV-Surgery, ARCTIC-HOI에서 zero-shot 성능 확인
- 손과 물체의 동적 움직임을 포함한 4D 구조 복원 가능
-

Qualitative Results

- 전체 장면의 3D 구조를 조밀하게 복원
- 손과 물체의 움직임을 시간적으로 보존
- 학습하지 않은 1인칭 영상에서도 안정적인 결과

Takeaway

EgoMono4D는 단안 1인칭 영상에서 depth와 camera motion을 결합해, 라벨 없이 빠르고 조밀한 4D pointcloud sequence를 생성

Thank you

Question & Disscusion

RILS LAB