Vision

RILS LAB

Robot Intelligence Learning & System

최성용

2026. 08. 14

01

VGGT: Visual Geometry Grounded Transformer

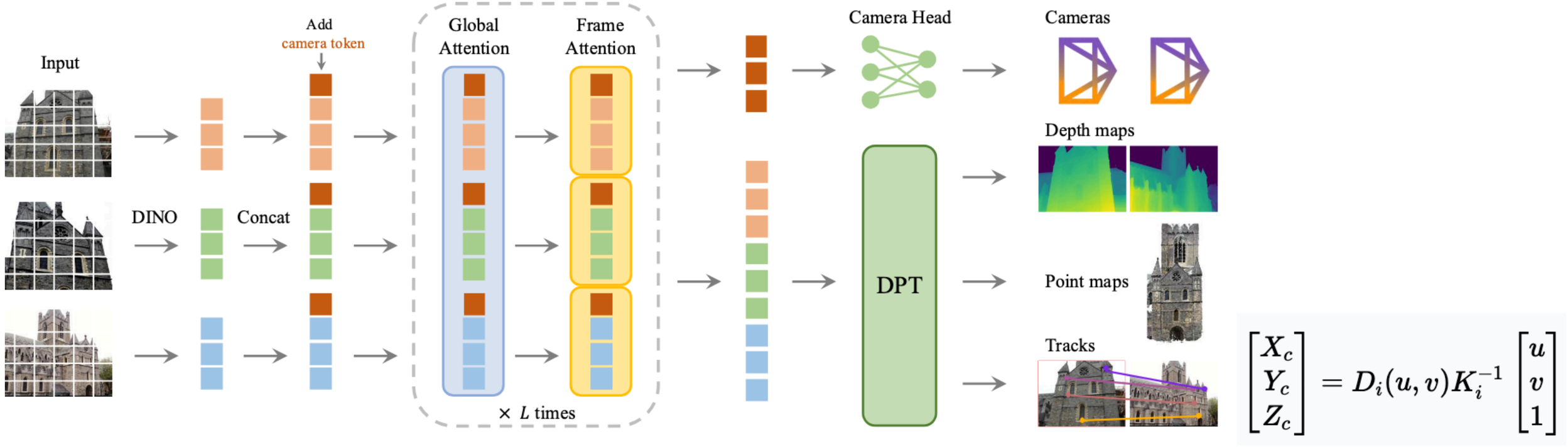
Jianyuan Wang, Minghao Chen, Nikita Karaev,
Andrea Vedaldi, Christian Rupprecht, David Novotny

CVPR 2025 Best Paper Award

Problem Setting

- 기존 SfM / MVS 방식은 matching 중심의 후처리가 필요
- 목표: 하나의 model로 camera parameter, depth map, point map, track을 예측
- 핵심 질문: 모델이 3D geometry 정보를 feed-forward로 후처리 없이 직접 복원 가능한가?

Solution: Methodology



$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = D_i(u,v)K_i^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

- 첫 번째 이미지는 전체 장면의 기준 좌표계로 사용
- Alternating-Attention: Global Attention-이미지 간 대응 관계 학습/ Frame Attention 세부적인 시각 구조 처리
- Camera Head: camera prameter 예측
- DPT: 저해상도 feature를 dense feature map으로 복원 + 각각 3x3 conv 적용 → 출력 Head
  - Depth map을 이용 → Point Map 생성
  - tracks: 별도 tracking 모듈 Co Tracker2 사용

Contribution

- cameras parameter, depth map, point maps, tracks을 하나의 모델로 통합 예측함
- 기존 DUSt3R, MASt3R는 2장만 입력 가능했던 것과 달리, 1장부터 수백 장까지 다양한 input image 지원
- camera pose estimation 및 multi-view depth estimation 등 다양한 3D task에서 SOTA 달성

02

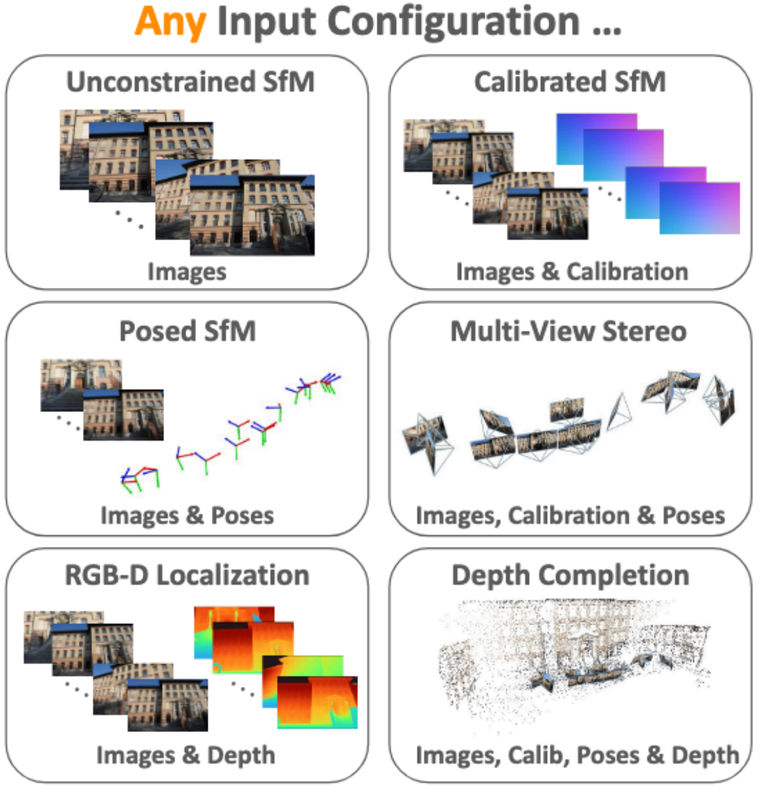
MapAnything: Universal Feed-Forward Metric 3D Reconstruction

Nikhil Keetha, Norman Müller, Johannes Schönberger,
Lorenzo Porzi, Yuchen Zhang, Tobias Fischer,
Sebastian Scherer, Peter Kontschieder

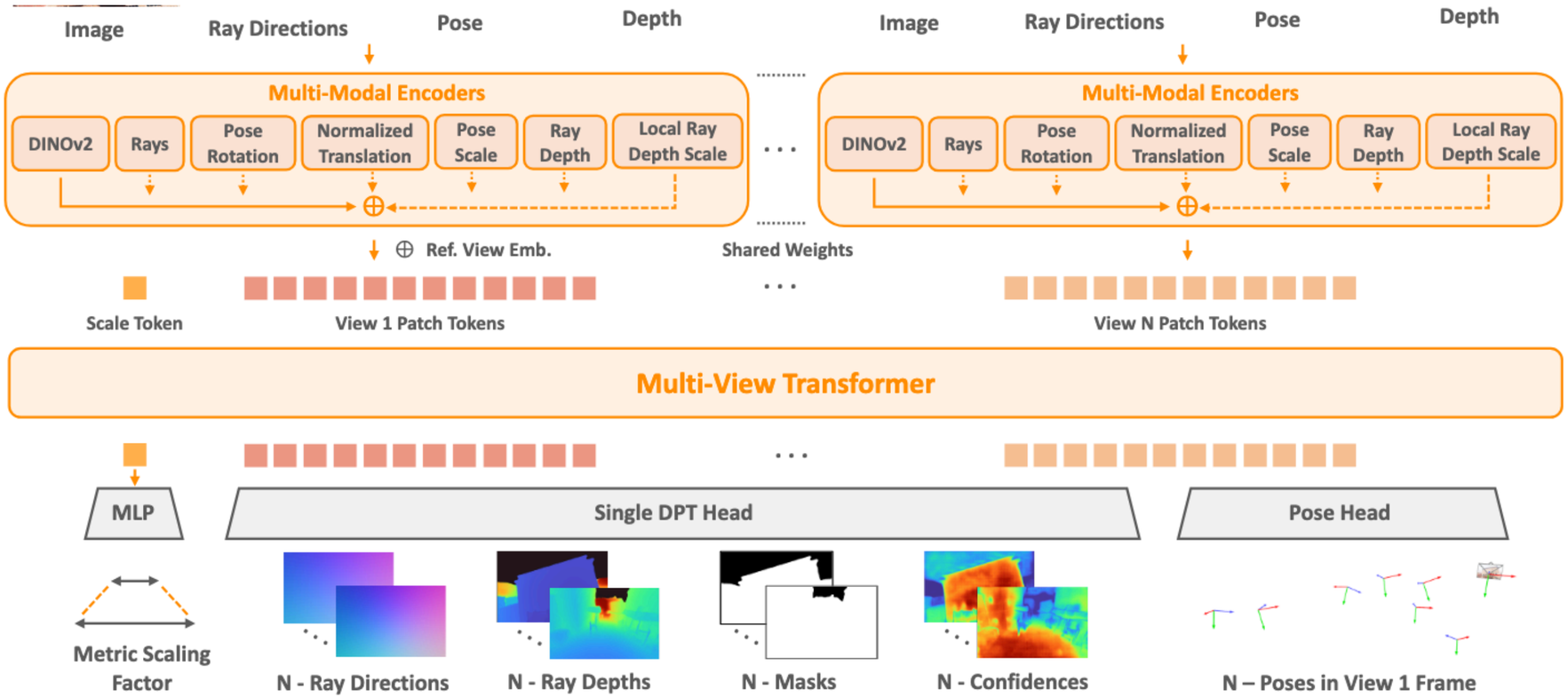
3DV 2026

Problem Setting

- 기존 image-only feed-forward 3D Foundation Model는 up-to-scale 한계가 있음
- 목표: local reconstruction 결과를 global한 metric 3D frame 으로 변환하는 것
- 입력: N개 RGB 이미지와 optional Geometry 입력을 함께 사용함
- 핵심 질문: 다양한 input 조건에서 metric 3D reconstruction을 예측할 수 있을까?



Solution: Methodology



$$\tilde{L}_i(u,v) = \tilde{R}_i(u,v) \cdot \tilde{D}_i(u,v)$$

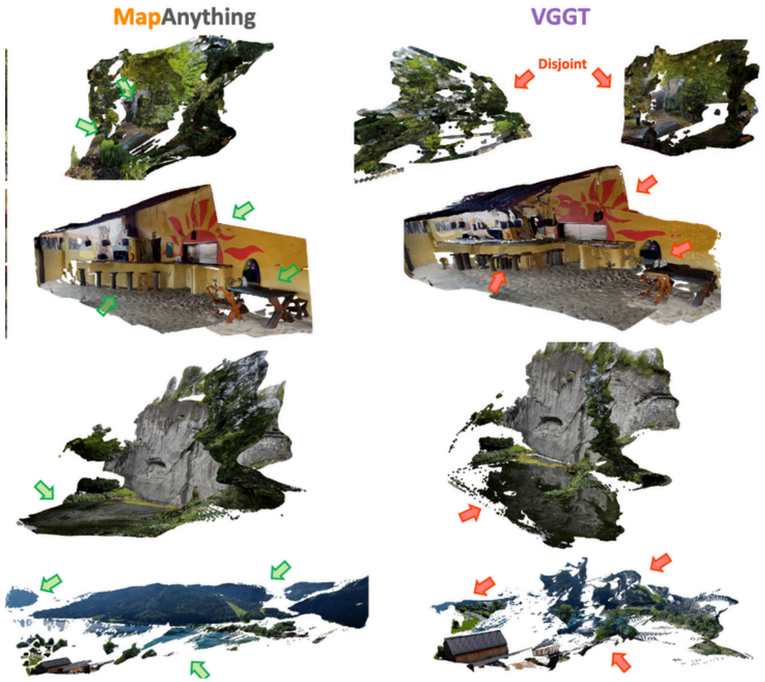
$$\tilde{X}_i = O_i \cdot \tilde{L}_i + \tilde{T}_i$$

$$X_i^{metric} = m \cdot \tilde{X}_1$$

- Multi Modal Encoder: input을 각 encoder로 처리 → 공통 latent space으로 변환 → patch token
  - 1번 view에는 **Reference view Embedding** 추가: 기준 좌표계
- Multi-View Transformer:
  - **scale token**: scene 전체에 공통 적용 → scene의 실제 metric scale 학습 → **metric scaling factor(m)**
  - patch token: 각 view의 국소적 정보
- Point map을 직접 예측하는 것이 아닌, **Ray map와 Ray Depth map을 통해 예측**함.

Contribution

- 범용 feed forward 방식의 3D 백본 네트워크를 제안함
- Factored Representation을 통해 local ray, ray depth, camera pose, metric scale을 분리하여 처리함
- Optional Geomeric 입력을 유연하게 활용할 수 있도록 설계



03

MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors

Riku Murai, Eric Dexheimer, Andrew J. Davison

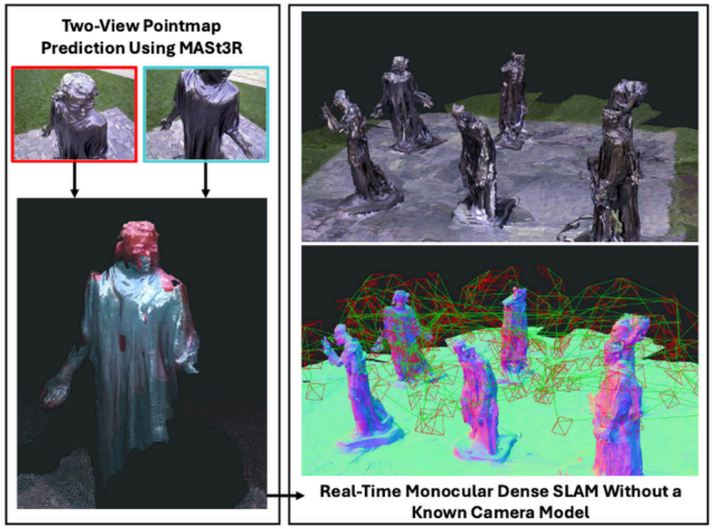
CVPR 2025 · Highlight

Problem Setting

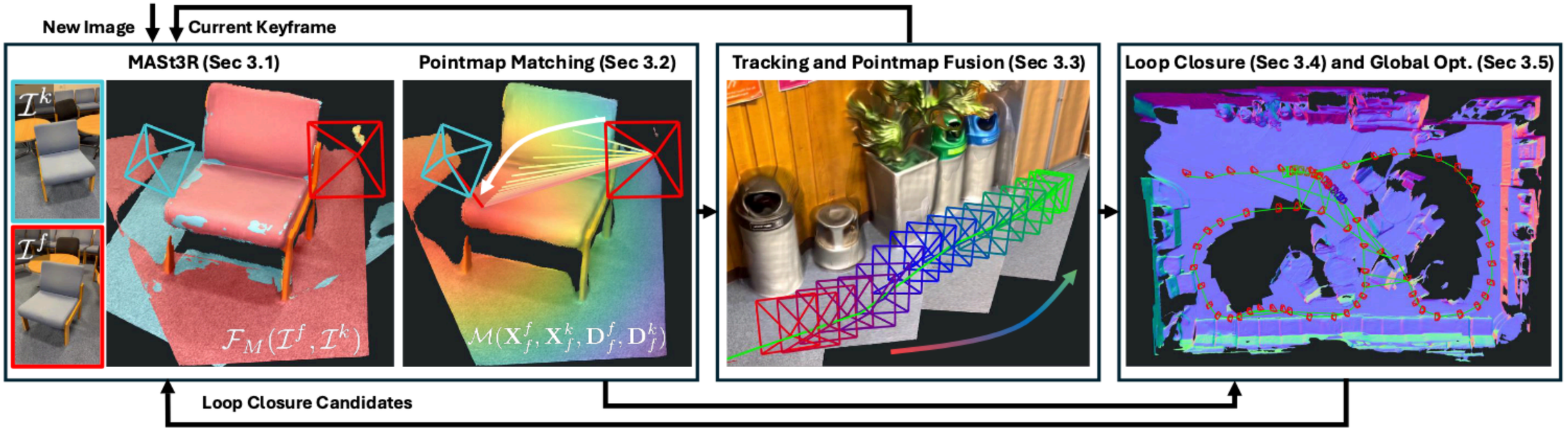
- 기존 SLAM은 calibration·pinhole camera model·sparse geometry 또는 무거운 최적화에 의존함

- 목표: camera tracking + globally consistent dense 3D map을 실시간 추정

- 핵심: MASt3R의 learned 3D reconstruction prior를 online SLAM으로 확장하는 것



Solution: Methodology



1. MASt3R: **Keyframe + New Image** → dense pointmaps,confidence, matching features 출력
2. Pointmap Matching: **iterative projection**(반복적으로 보정) 후, **local feature refinement**(대응점 미세 조정) 수행
3. Tracking & Fusion: Sim(3) + ray error를 이용해 pose 추정, tracking & pointmap을 confidence-weighted fusion함
4. loop closure edge & global optimisation: 현재 Keyframe과 과거 Keyframe 사이에 새로운 연결 추가 및 Keyframe pose와 pointmap 사이의 일관성을 개선

$$T = \begin{bmatrix} sR & t \\ 0 & 1 \end{bmatrix}$$

Contribution

- MASt3R 3D prior 기반 real-time dense monocular SLAM을 구현함
- pointmap matching을 약 2 ms로 줄여 dense matching의 실시간성을 확보
- globally consistent pose + dense geometry를 약 15 FPS로 추정함

04

Ov3R: Open-Vocabulary Semantic 3D Reconstruction from RGB Videos

Ziren Gong, Xiaohan Li, Fabio Tosi, Jiawei Han,
Stefano Mattoccia, Jianfei Cai, Matteo Poggi

CVPR 2026

Problem Setting

- Open Vocabulary Semantic(OVS)이란?

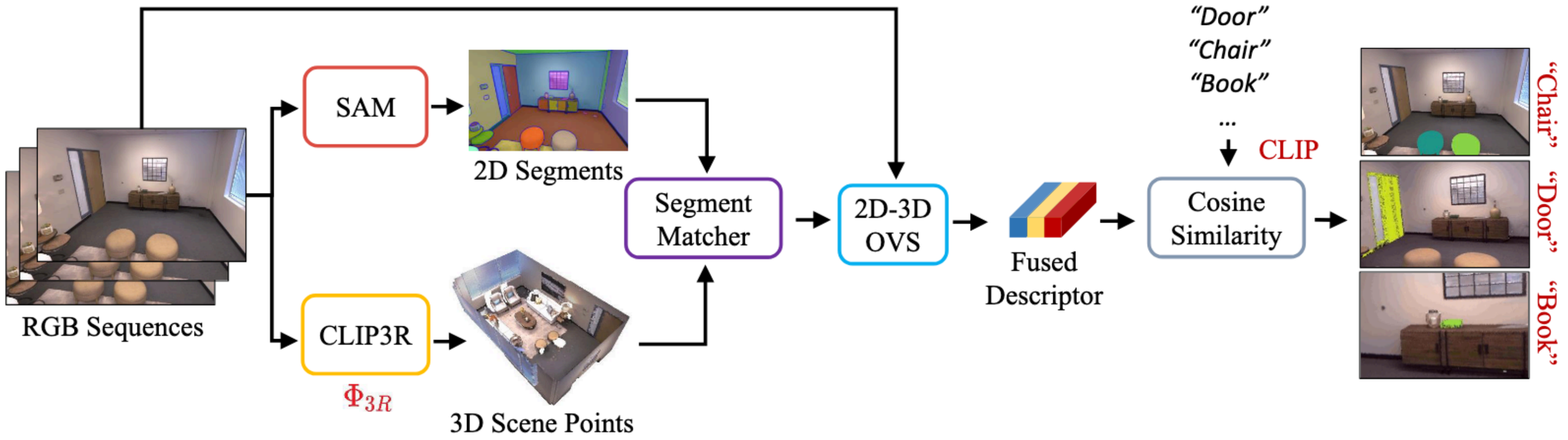
→ 미리 정의된 class 뿐만 아니라, 새로운 class도 text query를 통해 찾는 방법

- 기존에는 reconstruction과 semantic processing이 분리됨 → semantic-geometry alignment가 약함

- 목표: dense 3D reconstruction + open-vocabulary 3D semantics를 동시에 추정

- 핵심: RGB만으로 3D를 복원하면서, free text query로 3D 공간을 이해하는 것

Solution: Methodology



- **Segment Matcher**가 2D Segments와 3D Point 연결
- **2D–3D OVS**: 2D semantic 정보 + 3D geometric 정보를 입력 받아 fuse → **fused descriptor** 생성
- **CLIP Text Encoder**: "Door"와 같은 텍스트를 **CLIP text embedding**으로 변환
- CLIP text embedding과 Fused descriptor의 vision 정보의 cosine similarity 계산 → similarity가 가장 높은 text label을 해당 3D 영역에 매칭
- → open-vocabulary 3D segmentation을 수행

Contribution

- RGB-only 3D reconstruction과 open-vocabulary semantics를 하나의 framework로 통합
- CLIP semantics를 reconstruction 과정에 주입해, semantic–geometry alignment를 강화함
- SAM 기반 2D segment를 3D point map으로 대응시켜, instance-aware representation을 구성함

Thank you

Question & Disscusion

RILS LAB