DCHM: Depth-Consistent Human Modeling for Multiview Detection

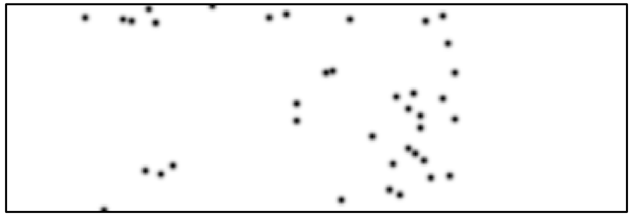
ICCV 2025

Jiahao Ma[1,2], Tianyu wang[1], Miaomiao Liu[1], David Ahmedt−Aristizabal[2] , Chuong Nguyen[2]
Australian National University[1], CSIRO DATA61[2]

Quick Preview - Research problem & Existing works

Research problem

Input: Multi-view RGB images



Output: Multi-view RGB images

Challenges:

- Sparse-view setting with limited overlapping
- Heavy occlusion in crowded scenes

Existing works

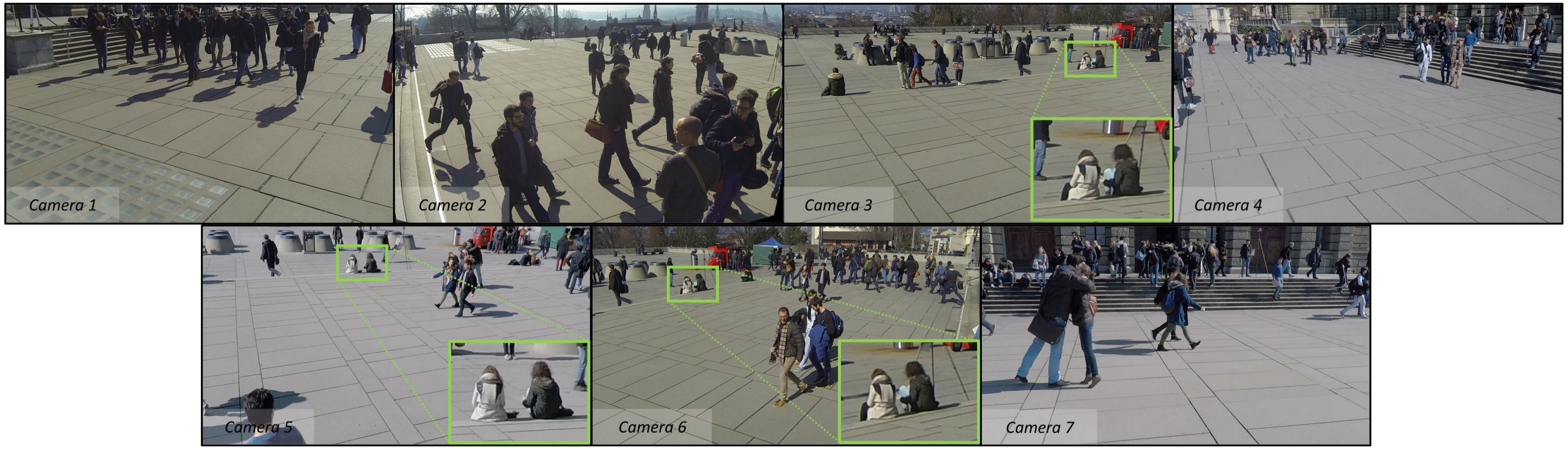
1. *Labeled-based methods* [1] [2] [3]
   - *Pros: High performance*
   - *Cons:*
     - *Dependence on cost labels*
     - *Poor robustness in diverse environments*

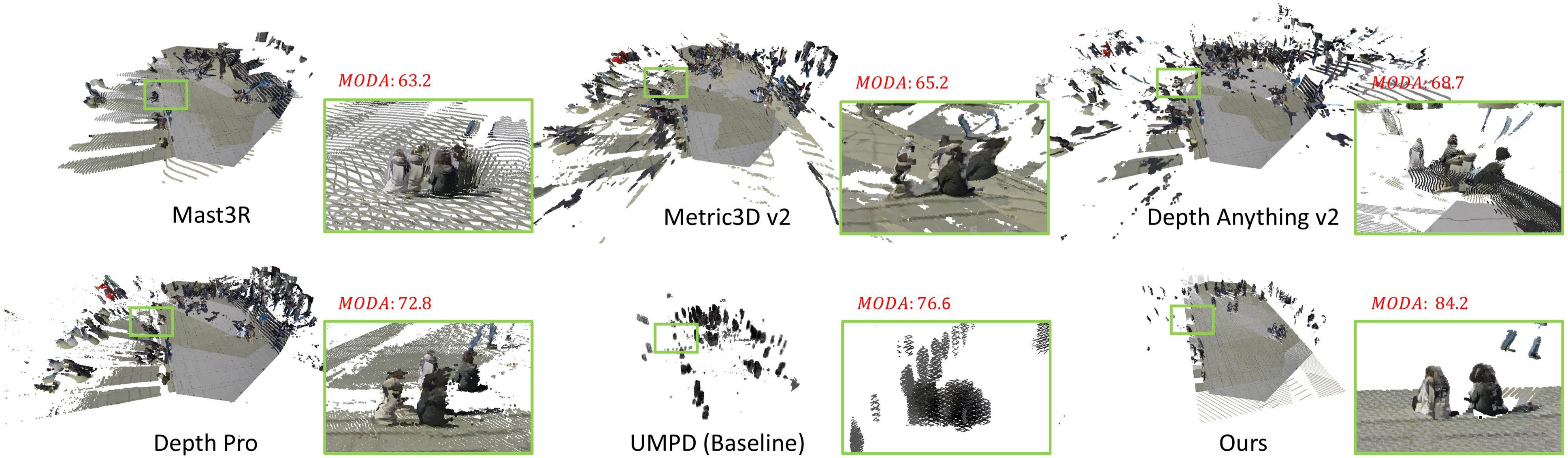
2. *Label-based methods* [4]
   - *Pros:*
     - *Do not require labels*
   - *Cons:*
     - *Low performance*

[1] Multi-view detection with feature perspective transformation. In ECCV 2020.
[2] Stacked homography transformations for multi-view pedestrian detection. In CVPR2021.
[3] Multiview detection with cardboard human modeling.. In ACCV2024.
[4] Unsupervised multi-view pedestrian detection. In ACMM2024.

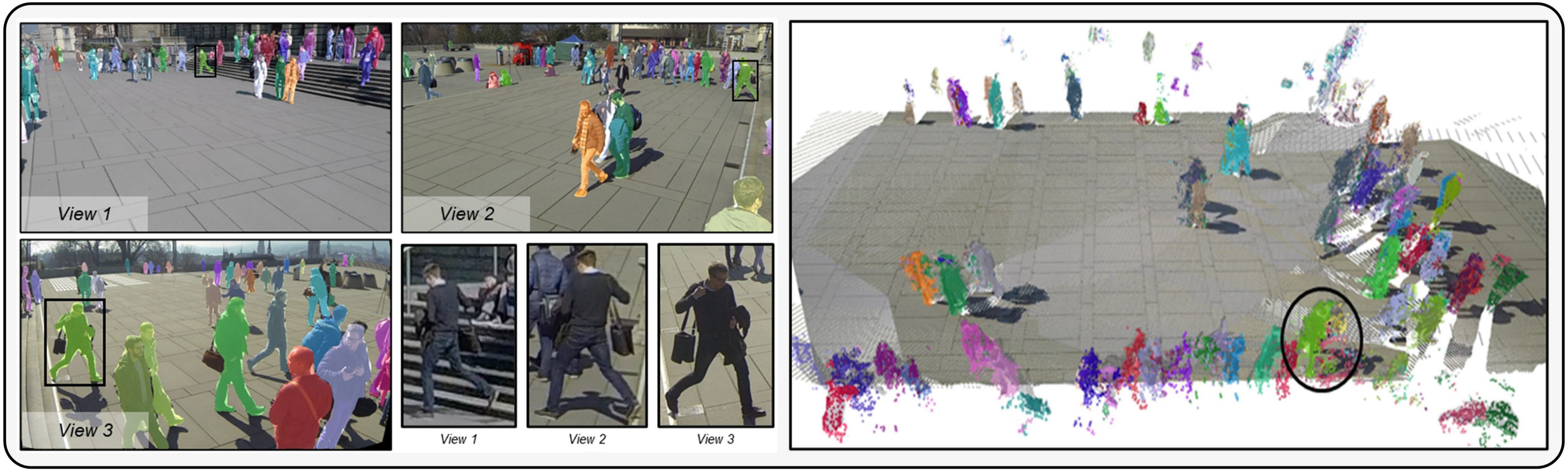
Quick Preview - Baselines

Mono-depth estimation:
lack of multi-view consistency



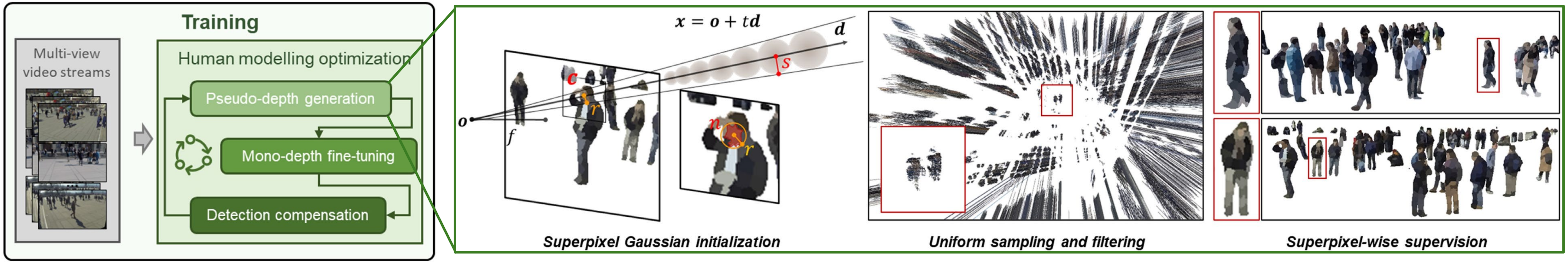


Method – Human modeling



Human modeling. Pedestrians are modeled as **segmented Gaussian primitives**, enabling robust multi-view fusion and detection even in *crowded* and *occluded* scenes.

Method – Training



Training

Multi-view video streams

Human modelling optimization

Pseudo-depth generation

Mono-depth fine-tuning

Detection compensation

Superpixel Gaussian initialization

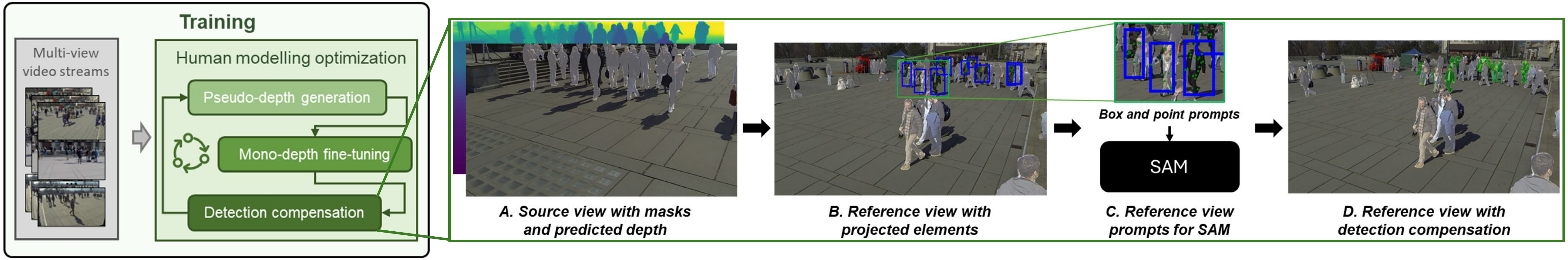
Uniform sampling and filtering

Superpixel-wise supervision

Training - We refine monocular depth using consistent pseudo-depth label across views.

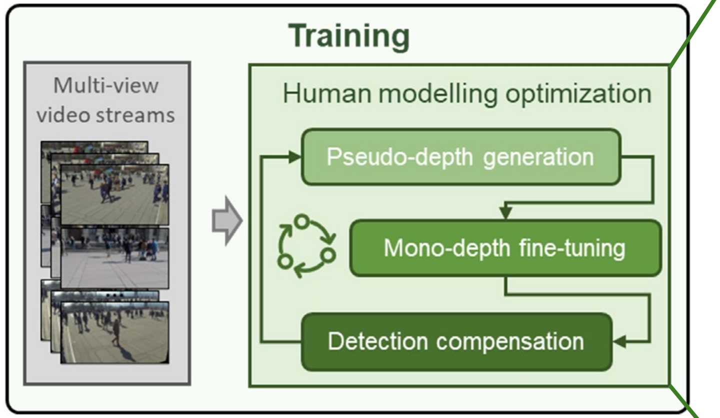
Consistent pseudo depth generation. We propose **superpixel-based initialization** method to allow GS optimization from sparse-view images.

Method – Training


Training
Multi-view video streams
Human modelling optimization
Pseudo-depth generation
Mono-depth fine-tuning
Detection compensation
Box and point prompts
SAM
A. Source view with masks and predicted depth
B. Reference view with projected elements
C. Reference view prompts for SAM
D. Reference view with detection compensation

Multi-view compensation for missed detection. We propose compensate the miss detection to generate better pseudo depth label and human mask for better optimization.

Method – Training



Training

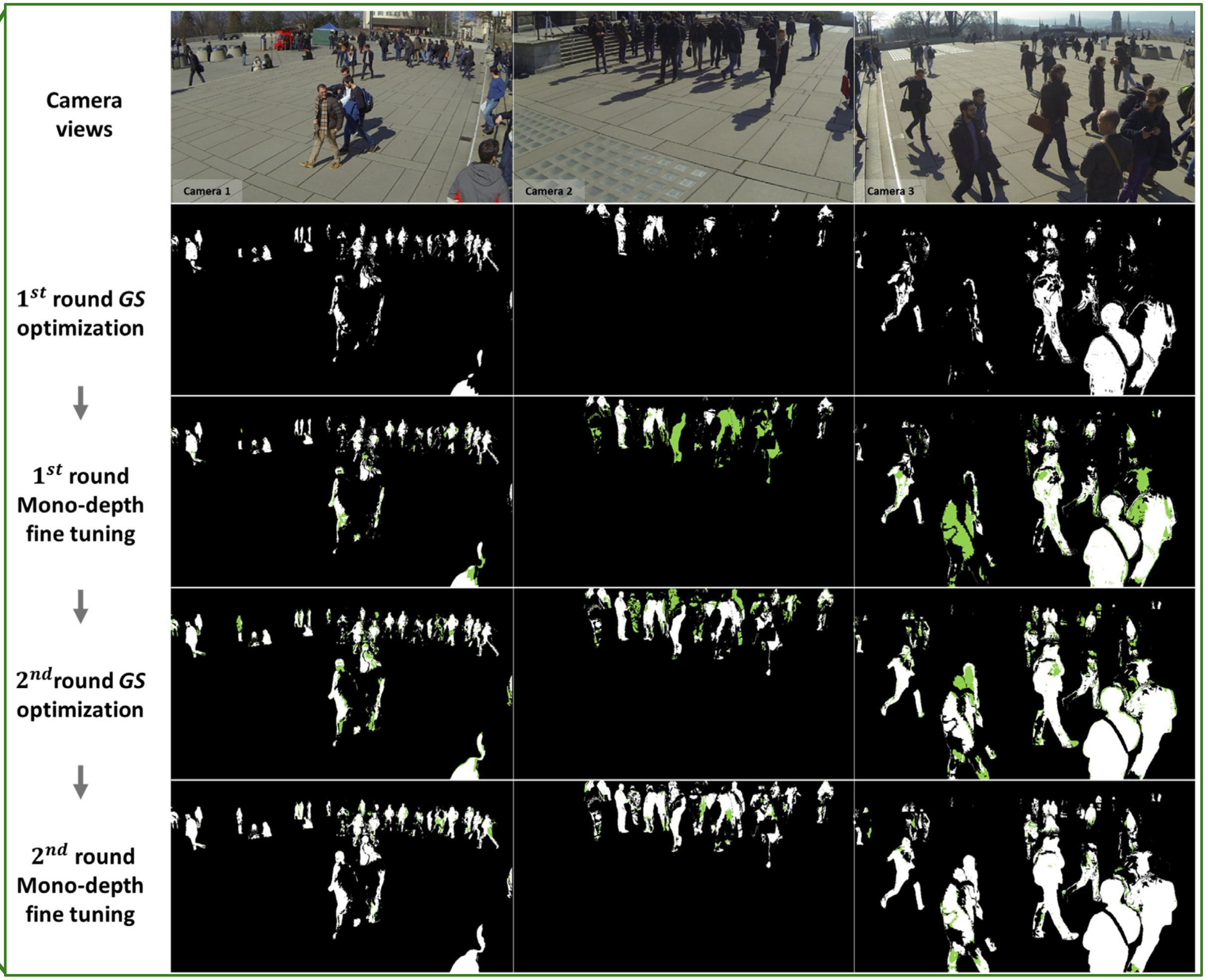
Multi-view video streams

Human modelling optimization

Pseudo-depth generation

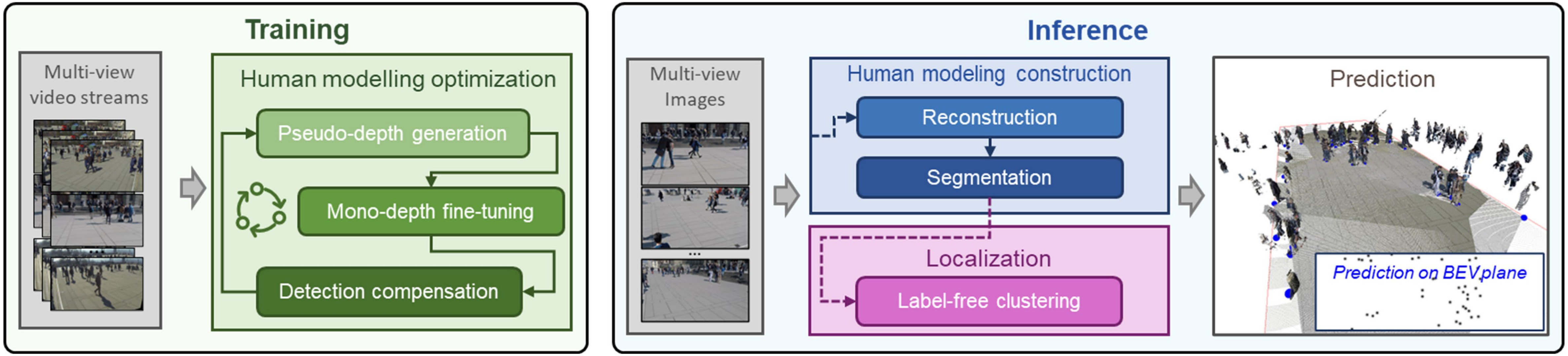
Mono-depth fine-tuning

Detection compensation


Camera views
Camera 1
Camera 2
Camera 3
1^{st} round *GS* optimization
1^{st} round Mono-depth fine tuning
2^{nd}round *GS* optimization
2^{nd} round Mono-depth fine tuning

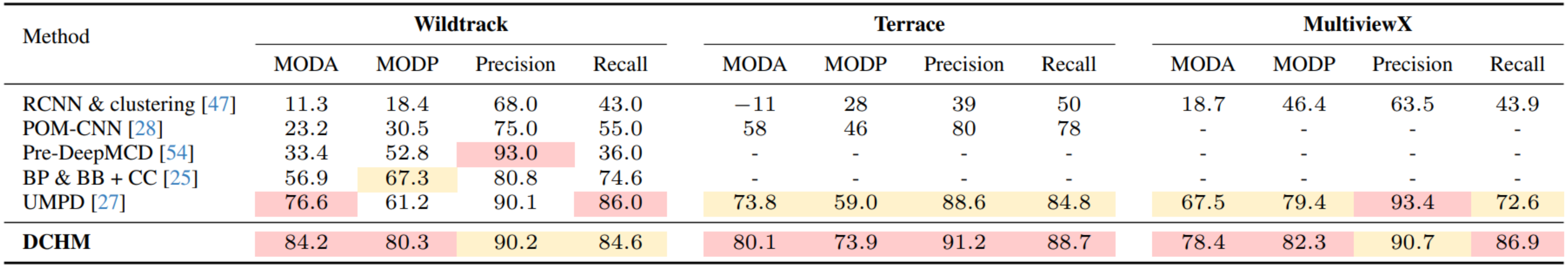
Iterative improvement. The pseudo-depth generation using GS, fine-tuning of mono-depth estimation, and multi-view detection compensation create an **iterative training loop**.

Method – Training ⇨ Inference pipeline



Framework Overview. We refine monocular depth using consistent pseudo-depth label across views. At inference, the depth-derived Gaussian primitives **are segmented and clustered in BEV** for pedestrian detection.

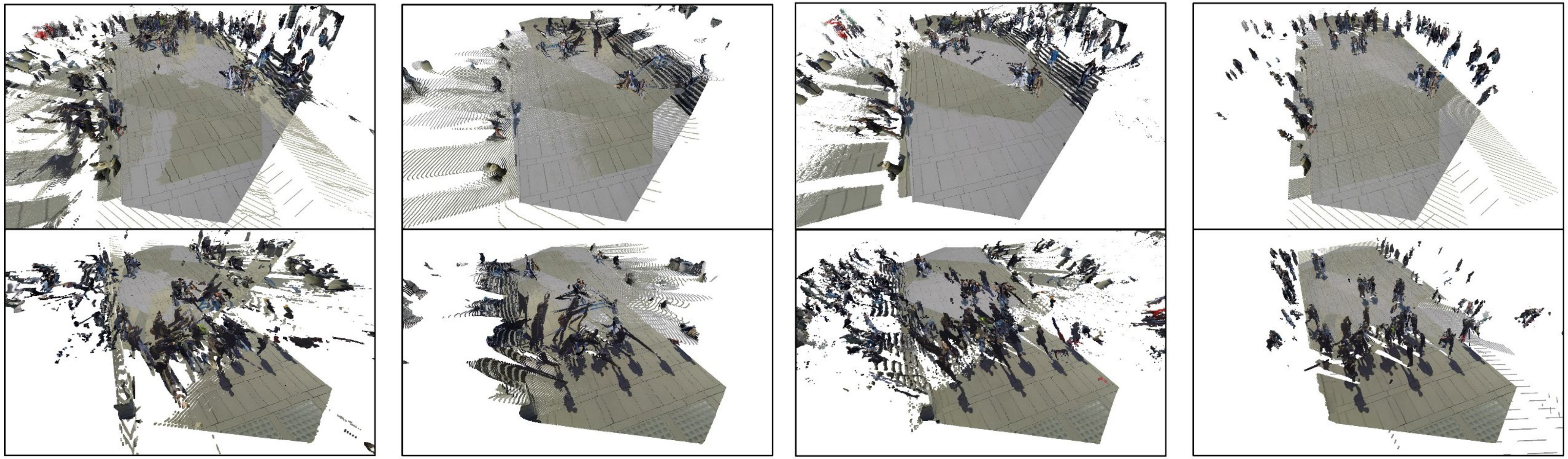
Results



Method	Wildtrack				Terrace				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
RCNN & clustering [47]	11.3	18.4	68.0	43.0	−11	28	39	50	18.7	46.4	63.5	43.9
POM-CNN [28]	23.2	30.5	75.0	55.0	58	46	80	78	-	-	-	-
Pre-DeepMCD [54]	33.4	52.8	93.0	36.0	-	-	-	-	-	-	-	-
BP & BB + CC [25]	56.9	67.3	80.8	74.6	-	-	-	-	-	-	-	-
UMPD [27]	76.6	61.2	90.1	86.0	73.8	59.0	88.6	84.8	67.5	79.4	93.4	72.6
DCHM	84.2	80.3	90.2	84.6	80.1	73.9	91.2	88.7	78.4	82.3	90.7	86.9

Achieve best or competitive performance on Wildtrack, Terrace and MultiviewX dataset.

Visual Comparison - Reconstruction



Metric3D v2

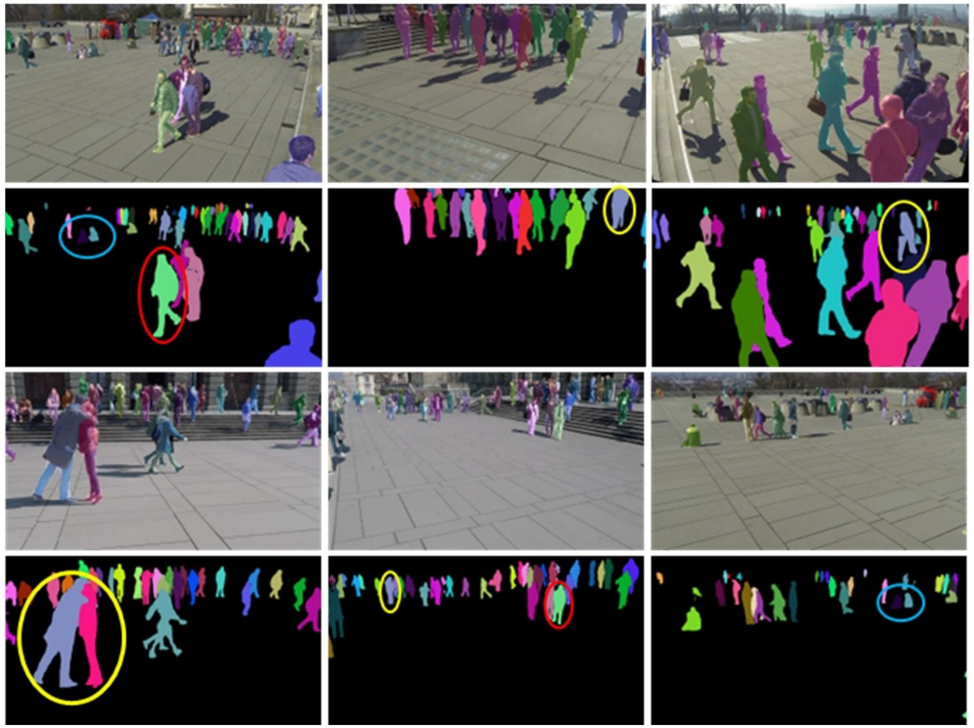
Mast3R

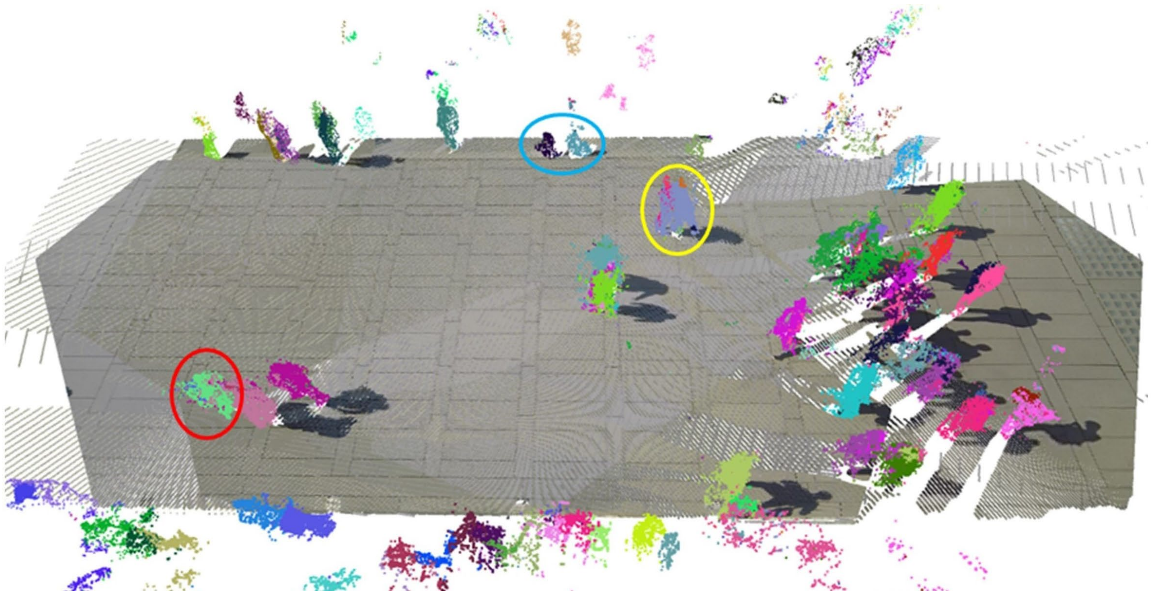
Depth Pro

Ours

Our method ("*Ours*") yields more **complete** and **accurate** 3D reconstructions than baselines, as shown in *front* and *back* views.

Visual Comparison - Segmentation





Pedestrians are clustered as **segmented Gaussians without labels**, with unique 3D IDs visualized as colour-consistent circles across 2D views.

DCHM: Depth-Consistent Human Modeling for Multiview Detection

Project page: https://jiahao-ma.github.io/DCHM/

GitHub

Scan me!