Spatial Visibility and Temporal Dynamics: Rethinking FoV Prediction in Adaptive PCV Streaming

*Chen Li, Tongyu Zong, Yueyu Hu, **Yong Liu,** and Yao Wang*

ACM Multimedia Systems Conference, April 1st, 2025

Outline

❑ Motivation: FoV adaptive 3D streaming

❑ CellSight -- visibility features

❑ CellSight -- prediction architecture

❑ Evaluation

❑ Conclusion & Future Work

Motivation: immersive 3D video

- ❑ Volumetric Capturing of Objects/Scenes
  - rich 3D information: geometry, color, texture
    - mesh, point cloud, 3D Gaussian splatting, etc.
  - immersive user viewing experience
    - 6 Degree of Freedom (DoF) viewing:
      position *(X, Y, Z)* and rotation *(Yaw, Pitch, Roll)*

Volumetric Capturing of Dancers @NYU

- ❑ Streaming Bandwidth Challenge:
  - A point cloud video consisting of 1M points per frame requires streaming bandwidth of more than 3.6Gbps without compression and 120 Mbps even with lossy compression.

Ohji Nakagami, Sebastien Lasserre, Sugio Toshiyasu, and Marius Preda. 2023. White paper on G-PCC. In ISO/IEC JTC 1/SC 29/AG 03 N0111.
https://www.mpeg.org/wp-content/uploads/mpeg_meetings/142_Antalya/w22804.zip

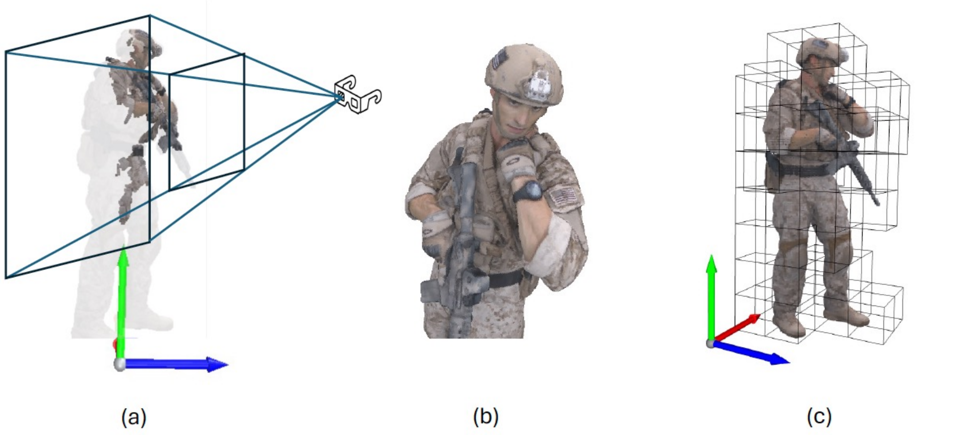
Motivation: FoV adaptive streaming

❑ Finite Field-of-View (time-varying)

❑ 3D Cell-based encoding/delivery

❑ Allocating more bits to more “visible” cells

- higher bitrates for cells inside predicted FoV (similar to 360)
- within FoV, lower bitrates for farther away cells, less intelligible (new to 3D)



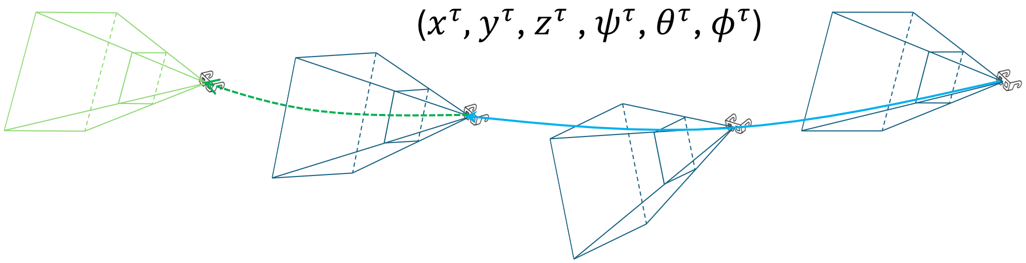
visible points   actual view   cell encoding

<u>*Streaming efficiency hinges on accurate cell visibility prediction:*</u>

1) *how to quantify each cell’s “contribution” to viewing quality?*
2) *how to predict them in near future?*

Motivation: how to accurately predict FoV?

Trajectory-based methods predict future 6-DoF coordinates as time-series, then calculate point/cell visibility using predicted values.



$$(x^\tau, y^\tau, z^\tau, \psi^\tau, \theta^\tau, \phi^\tau)$$

1) 6DoF prediction is much hard than 3DoF in 360-degree video. Small errors in each 6DoF coordinate may translate into large errors in visible points prediction due to **error amplification**.
2) Trajectory-based viewport prediction alone does not consider **the impact of video content**.

Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware mobile volumetric video streaming. In Proceedings of the 26th annual international conference on mobile computing and networking. 1–13.
Junhua Liu, Boxiang Zhu, Fangxin Wang, Yili Jin, Wenyi Zhang, Zihan Xu, and Shuguang Cui. 2023. Cav3: Cache-assisted viewport adaptive volumetric video streaming. In 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 173–183

LSTM/MLP for 200ms

Considering content but it is still trajectory-based

LLM Intelligence on FoV Prediction

Prompt: *how to predict FoV in point cloud video streaming?*

Learning from Historical Data: Use past frames of the point cloud stream to train a model that predicts future FoV based on the current sensor position and motion.
Using Recurrent Neural Networks (RNNs): RNNs, particularly Long Short-Term Memory (LSTM) networks, are good for modeling temporal sequences. You can train an LSTM to predict the next FoV based on previous frames and sensor motion.
Reinforcement Learning: In autonomous systems, reinforcement learning can be used to predict FoV based on the sensor's interaction with the environment, continuously learning and optimizing the field of view prediction.

(A) Traditional Methods (Linear Extrapolation)

•Use **velocity and acceleration models** to predict future head movements.....

(B) Machine Learning-Based Prediction

•**LSTM/RNN Networks**: Model temporal dependencies in head movement.

•**Transformer Models**: Capture long-range dependencies in viewing behavior.

•**Hybrid Models**: Combine head pose + eye-tracking for higher accuracy.

(C) Saliency-Based Prediction

Uses **3D saliency maps** derived from point cloud features (color, geometry, motion).

(D) Reinforcement Learning (RL)

•Train an RL agent to learn optimal FoV prediction policies based on user interaction feedback.

Outsmart LLMs: *why not **directly** predicting cell visibility?*

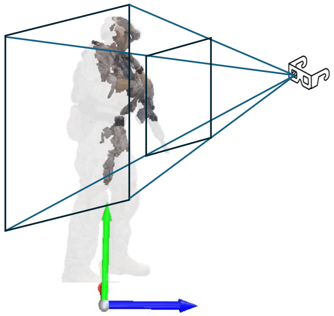
- ✓ Avoiding error amplification when translating from 6-DoF coordinates to cell visibility
- ✓ exploiting spatial/temporal locality for fine-grained predictions
  - **spatial:** if one cell is 'inside FoV', it's neighbors are more likely 'inside FoV';
  - **temporal:** if one cell is 'moving out of FoV' now, it's more likely this cell will have low visibility later;
  - **continuous** variations instead of binary predictions.
- ✓ organically integrating cell-based features
  - distance to viewpoint,
  - point density within a cell,
  - color, texture, motion……

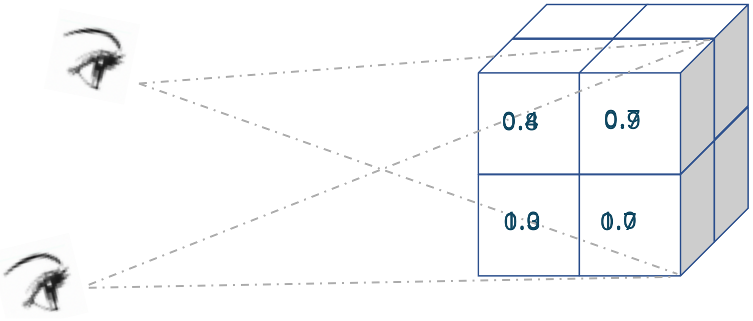
Outline

- ❑ Motivation: FoV adaptive 3D streaming
- ❑ CellSight -- visibility features
- ❑ CellSight -- prediction architecture
- ❑ Evaluation
- ❑ Conclusion & Future Work

Visibility Feature 1: viewport overlap ratio

How much volume of each cell is inside the viewport? (continuous number from 0 to 1)





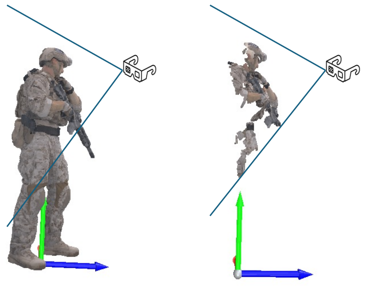
$$F^{\tau} = \left\{ f_i^{\tau} = \frac{K_i}{N_i}, i \in C \right\}$$

Instead of calculating the geometry directly, we uniformly scatter points in the whole space and calculate number of points inside viewport. It can be executed on all cell parallelly to quickly get the estimate features for all cells.

Visibility Feature 2: Occlusion-aware Visibility

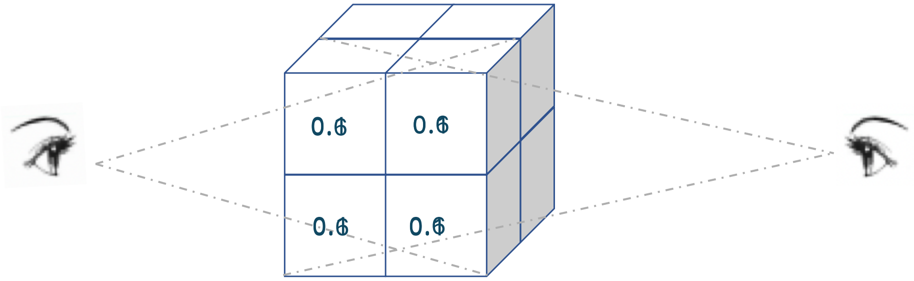
- ❑ Points may be occluded by other points in 3D Space
- ❑ Cell visibility taking into account occlusion
  - percentage of visible points after occlusion in each cell

$$V^T = \{\frac{v_i^T}{N_i}, i \in C\}.$$



(1)

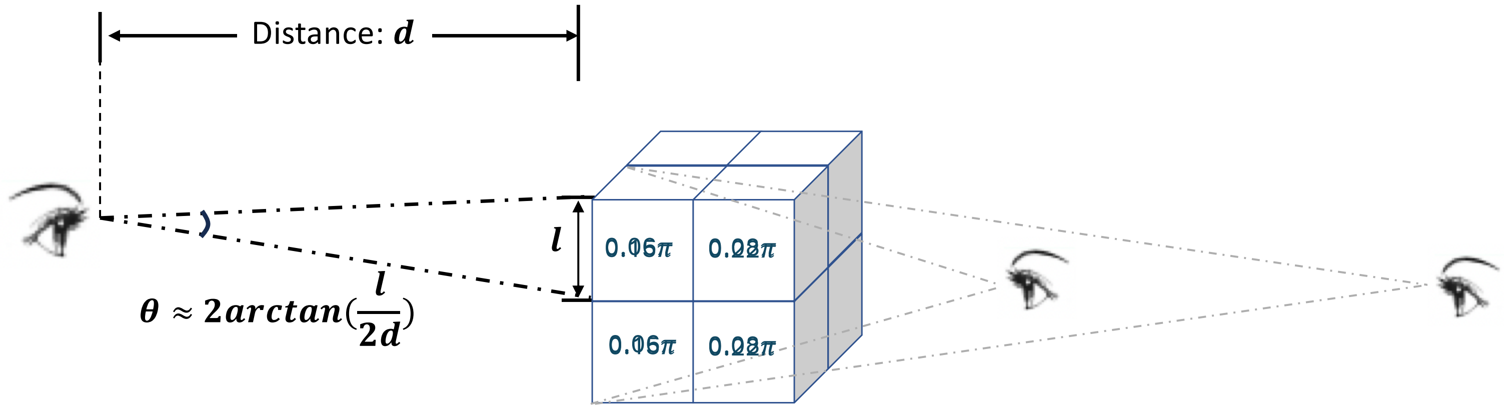
(2)



Hidden Point Removal (HPR) is time consuming (Katz, Tal, and Basri 2007).
Cell-based occlusion estimation method introduce significant quantization errors (Han, Liu, and Qian 2020)
We down sample the original point cloud video and perform HPR.

Visibility Feature 3: Angular Span/Visible Angular Span

- ❑ Cell far away from viewport contributes to a small angular span, viewing quality contribution saturates early at low bitrate.


Distance: $\boldsymbol{d}$
$\boldsymbol{l}$
$\boldsymbol{\theta \approx 2arctan(\frac{l}{2d})}$
0.06π
0.08π
0.06π
0.08π

$$A^{\tau} = \{\alpha_i^{\tau} = f_i^{\tau} * \theta_i^{\tau}, i \in C\}$$

f_i^{τ}: viewport overlap ratio for cell i at time τ

Cell Angular Span feature

$$B^{\tau} = \{\beta_i^{\tau} = v_i^{\tau} * \theta_i^{\tau}, i \in C\}$$

v_i^{τ}: occlusion-aware visibility for cell i at time τ

Visible Cell Angular Span feature

Outline

❑ Motivation: FoV adaptive 3D streaming

❑ CellSight -- visibility features

❑ CellSight -- prediction architecture

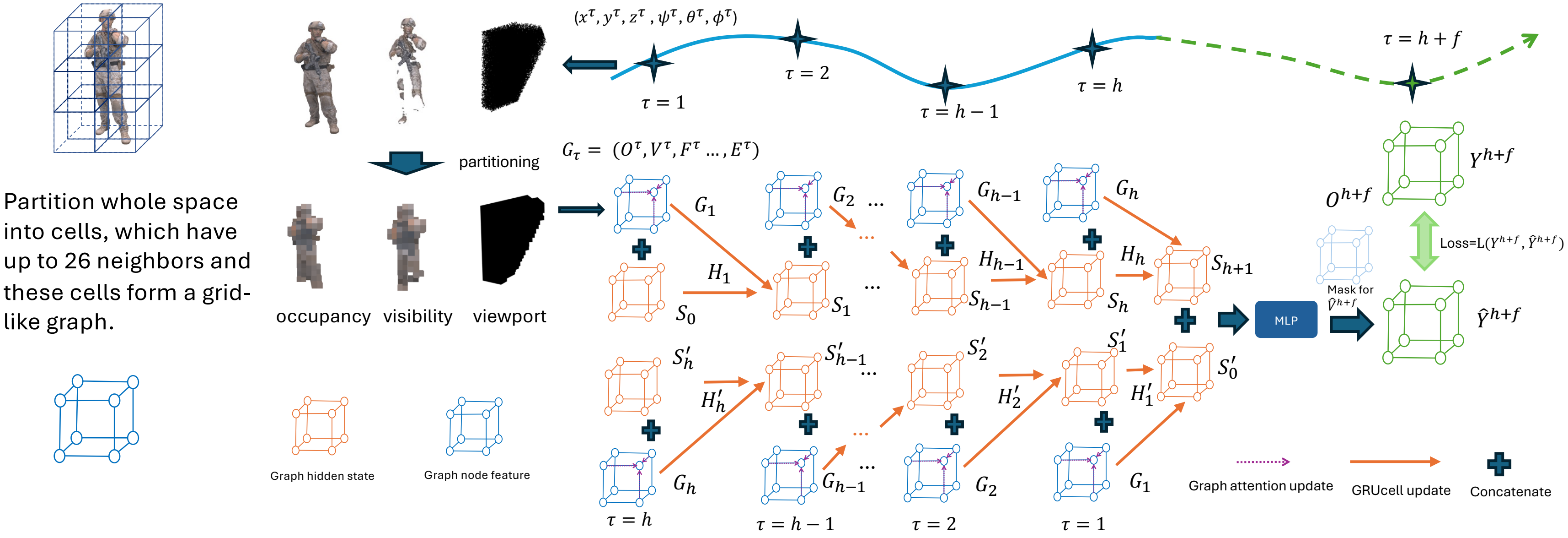
❑ Evaluation

❑ Conclusion & Future Work

CellSight: overview

Input-- history cell features: visibility, occupancy, cell center, viewing distance, etc.

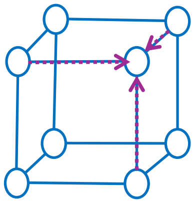
Output-- cell visibility in near future, e.g., 33ms ~ 5 sec.



Partition whole space into cells, which have up to 26 neighbors and these cells form a grid-like graph.

TransGraph/GRU Modules

- TransGraph: Transformer-based Graph Neural Network to capture spatial correlation between neighboring cells



$$\hat{H}_\tau = TransGraph(H_\tau)$$

GRU hidden state

$$H_\tau = \{h_i^\tau, i \in C\} = S_\tau \oplus G_\tau$$

raw cell features

$$q_{c,i}^\tau = W_{c,q} h_i^\tau + b_{c,q}$$

$$k_{c,j}^\tau = W_{c,k} h_j^\tau + b_{c,k}$$

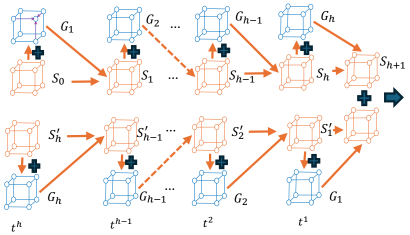
$$\alpha_{c,ij}^\tau = \frac{\langle q_{c,i}^\tau, k_{c,j}^\tau \rangle}{\sum_{u \in \mathcal{N}(i)} \langle q_{c,i}^\tau, k_{c,u}^\tau \rangle}$$

$$v_{c,j}^\tau = W_{c,v} h_j^\tau + b_{c,v}$$

$$\hat{h}_i^\tau = \sum_{j \in \mathcal{N}(i)} \alpha_{c,ij}^\tau v_{c,j}^\tau$$

Graph attention module over neighbors

- Bi-direction GRU: temporal dynamics of cell features



$$S_{t+1} = GRU_f(\hat{H}_t, G_t)$$

$$S'_{t-1} = GRU_r(\hat{H}'_t, G_t)$$

Graph output state

[14]Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2020). Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*.
[15] He, Hangtao, Linyu Su, and Kejiang Ye. "GraphGRU: A graph neural network model for resource prediction in microservice cluster." 2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2023.

Outline

❑ Motivation: FoV adaptive 3D streaming

❑ CellSight -- visibility features

❑ CellSight -- prediction architecture

❑ Evaluation

❑ Conclusion & Future Work

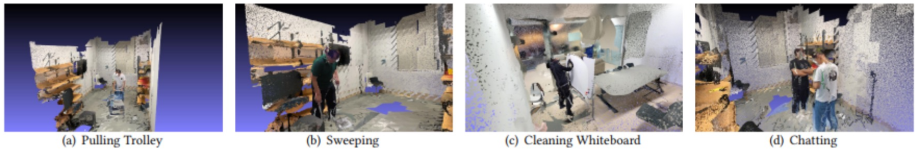
Datasets

Two public point cloud video datasets:
8i, four videos [16]
Full scene volumetric video dataset(FSVVD), four videos[17]

Users' 6DoF viewing navigation trajectory data
26 users, more than 40k frames[18]
12 users, more than 50k frames[19]

8i data screenshot

First three videos for training and last videos for testing and validation



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

FSVVD data screenshot

[16]Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 2017. 8i voxelized full bodies-a voxelized point cloud dataset. ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006 7, 8 (2017), 11.
[17]Kaiyuan Hu, Yili Jin, Haowen Yang, Junhua Liu, and FangxinWang. 2023. FSVVD: A dataset of full scene volumetric video. In Proceedings of the 14th Conference on ACM Multimedia Systems. 410–415.
[18] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User centered adaptive streaming of dynamic point clouds with low complexity tiling. In Proceedings of the 28th ACM international conference on multimedia. 3669–3677.
]19]Kaiyuan Hu, Haowen Yang, Yili Jin, Junhua Liu, Yongting Chen, Miao Zhang, and FangxinWang. 2023. Understanding user behavior in volumetric video watching: Dataset, analysis and prediction. In Proceedings of the 31st ACM International Conference on Multimedia. 1108–1116.

Baselines and Metrics

- **Linear Regression (LR):** A linear model predicts each of the 6DoF coordinate using a linear combination of the values on the same coordinate over a history window.
- **Truncated linear regression (TLR):**This approach uses the last monotonically increasing or decreasing part of the history window to linearly extrapolate the future value. It has shown a great performance in sequence prediction, in particular short-term FoV prediction[12].
- **Mutli-task Multilayer Perceptron (MLP):** Similar to the MLP model in [13][14]
- **Mutli-task LSTM(LSTM) [13][14]:** a two-layer LSTM model with 60 neurons per layer to predict future FoV coordinates based on historical data, predicting all coordinates simultaneously

LR and TLR predict all coordinates individually. MLP and LSTM predict them all together.

Visibility MSE over all cells and R^2 to assess the MSE relative to the variance of the ground truth:

$$R^2 = 1 - \frac{\text{MSE}}{\text{GT variance}} = 1 - \frac{\sum_i \sum_f (y_{if} - \hat{y}_{if})^2}{\sum_i \sum_f (y_{if} - \bar{y})^2}$$

[12]Chenge Li, Weixi Zhang, Yong Liu, and Yao Wang. 2019. Very long term field of view prediction for 360-degree video streaming. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 297–302.
[13]Xueshi Hou and Sujit Dey. 2020. Motion Prediction and Pre-Rendering at the Edge to Enable Ultra-Low Latency Mobile 6DoF Experiences. IEEE Open Journal of the Communications Society 1 (2020), 1674–1690.
[14]Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware mobile volumetric video streaming. In Proceedings of the 26th annual international conference on mobile computing and networking. 1–13.

Viewport Overlap Ratio Prediction on 8i

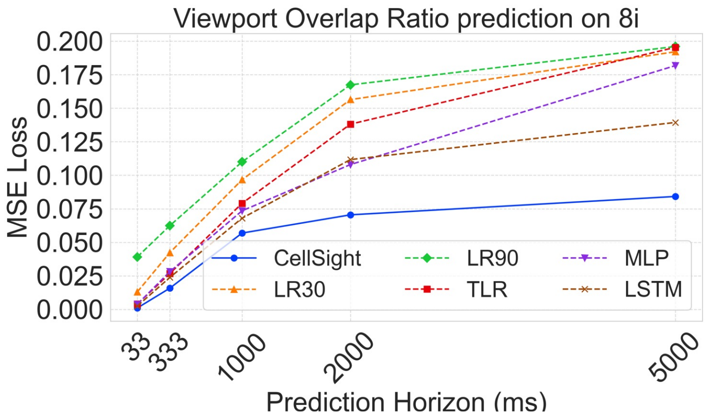
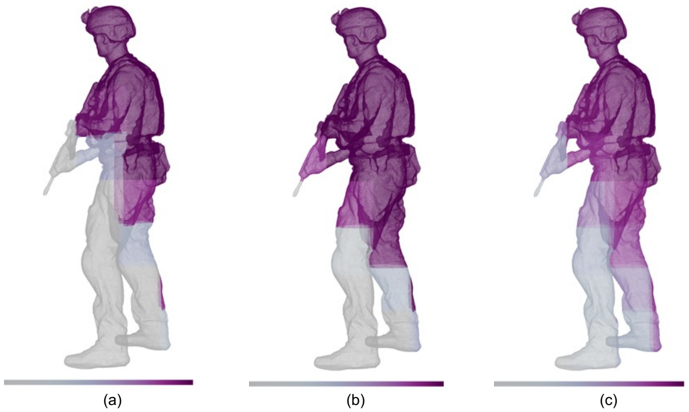


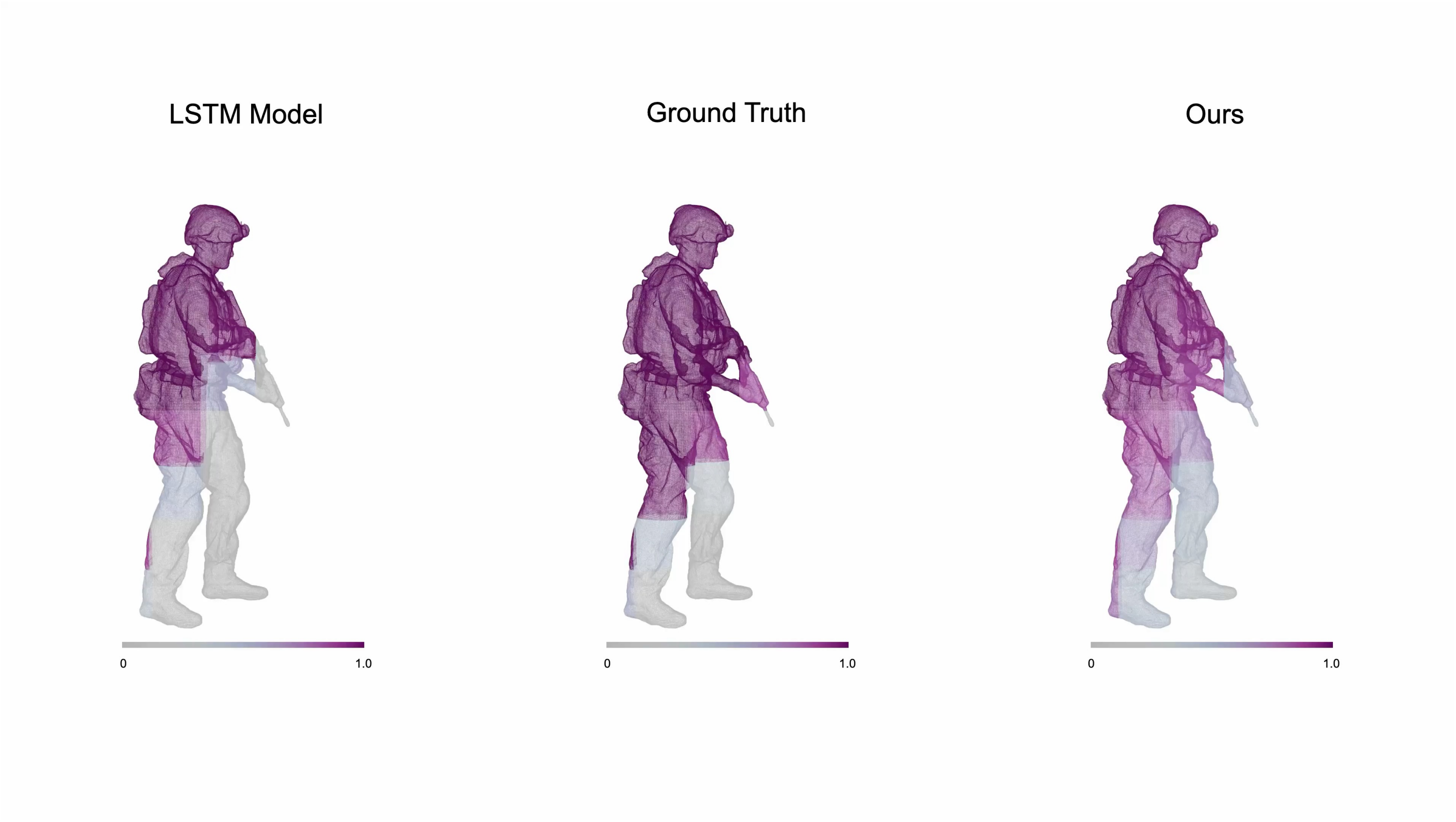
Table 2: Viewport Overlap Ratio R^2 Scores on 8i

Time (ms)	LR30	LR90	TLR	MLP	LSTM	CellSight
33	0.941	0.824	0.982	0.983	0.989	**0.995**
333	0.809	0.718	0.879	0.873	0.893	**0.928**
1000	0.563	0.504	0.644	0.669	0.693	**0.743**
2000	0.294	0.246	0.378	0.514	0.497	**0.682**
5000	0.135	0.114	0.116	0.178	0.370	**0.618**



Cell viewport overlap ratio prediction results on the 8i dataset, where the prediction horizon is 2000ms. Visual comparison of predicted viewport using (a) LSTM model, (b) Ground Truth, and (c) CellSight. The color represents the prediction confidence, transitioning from dark (low confidence) to bright (high confidence).

3D Visualization for One Frame



Occlusion-aware Visibility Prediction on FSVVD

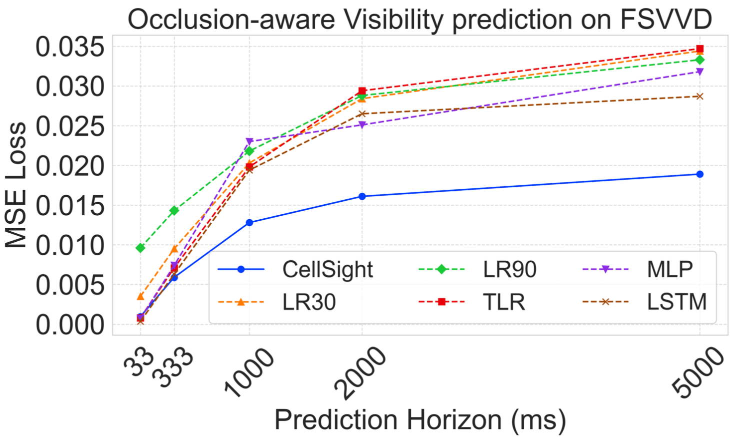
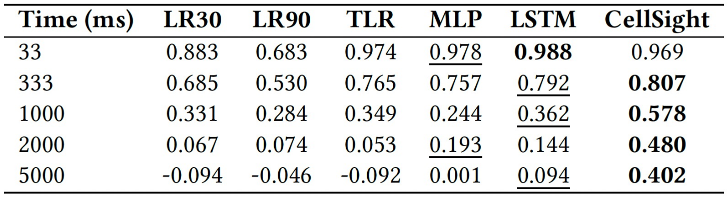


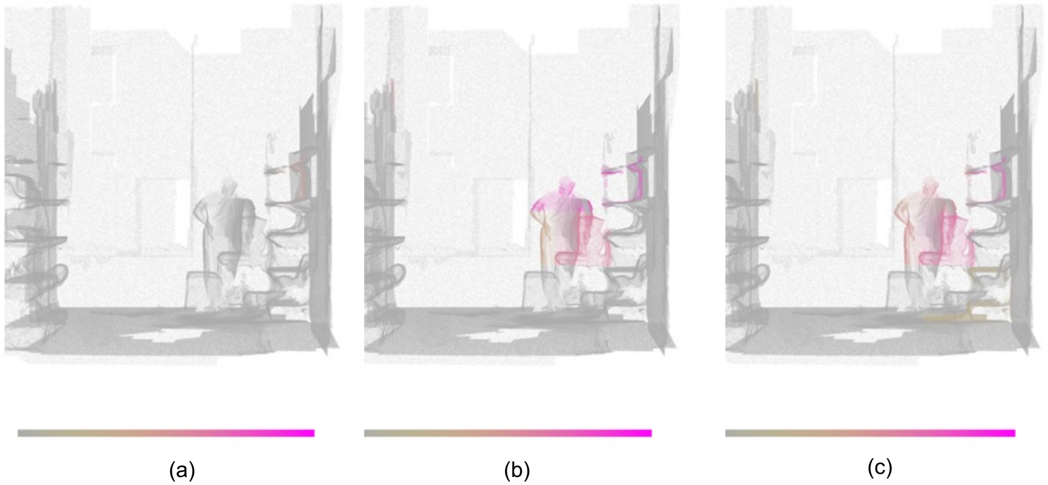
Table 8: Occlusion-aware Visibility R^2 Scores on FSVVD



Time (ms)	LR30	LR90	TLR	MLP	LSTM	CellSight
33	0.883	0.683	0.974	0.978	**0.988**	0.969
333	0.685	0.530	0.765	0.757	0.792	**0.807**
1000	0.331	0.284	0.349	0.244	0.362	**0.578**
2000	0.067	0.074	0.053	0.193	0.144	**0.480**
5000	-0.094	-0.046	-0.092	0.001	0.094	**0.402**

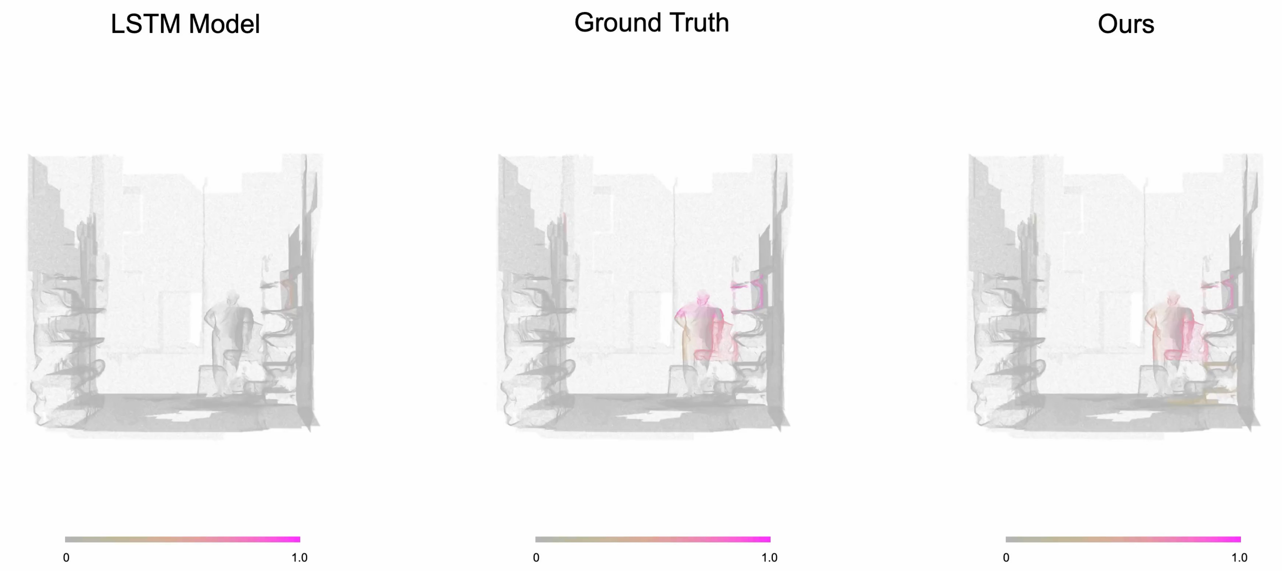
Since the user’s trajectory data is more dynamic in FSVVD dataset, some methods even produce negative R^2 scores at long time horizons, indicating poor performance. But our model still maintains positive scores, demonstrating its robustness and predictive capability even under challenging conditions.

Visible Angular Span on FSVVD



Visible angular span prediction for each cell on the FSVVD dataset at one frame, where the prediction horizon is 2000ms. (a) LSTM model, (b) Ground Truth, and (c) CellSight. The color transition from gray to red corresponds to visible angular span from small to large. The prediction by our approach closely matches the ground-truth, whereas LSTM produces significantly different results due to error amplification in the degrees of freedom, leading to substantial discrepancies in the final visible cell angular spans. R^2 = 0.54

3D Visualization

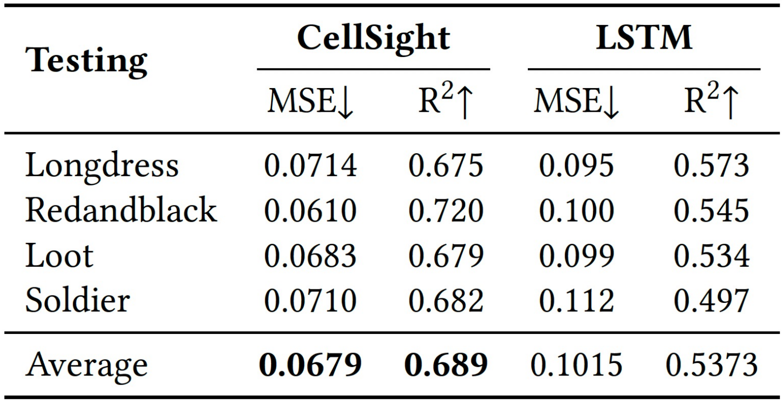


LSTM Model

Ground Truth

Ours

Cross Validation



Testing	CellSight		LSTM	
	MSE↓	R^2↑	MSE↓	R^2↑
Longdress	0.0714	0.675	0.095	0.573
Redandblack	0.0610	0.720	0.100	0.545
Loot	0.0683	0.679	0.099	0.534
Soldier	0.0710	0.682	0.112	0.497
Average	**0.0679**	**0.689**	0.1015	0.5373

Evaluate model generalization by rotating testing/validation among four PCVs on 8i.

CellSight has lower average MSE and higher R^2 compared with LSTM

CellSight shows good generalization.

Conclusion & Future Work

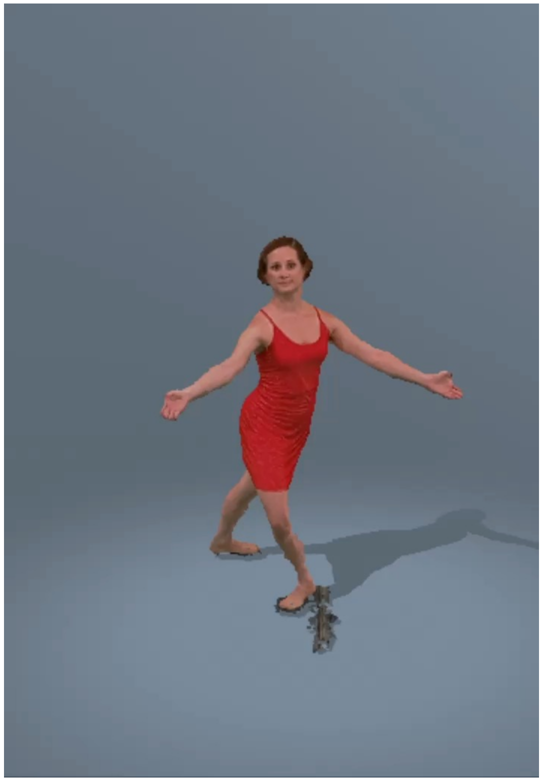
CellSight:

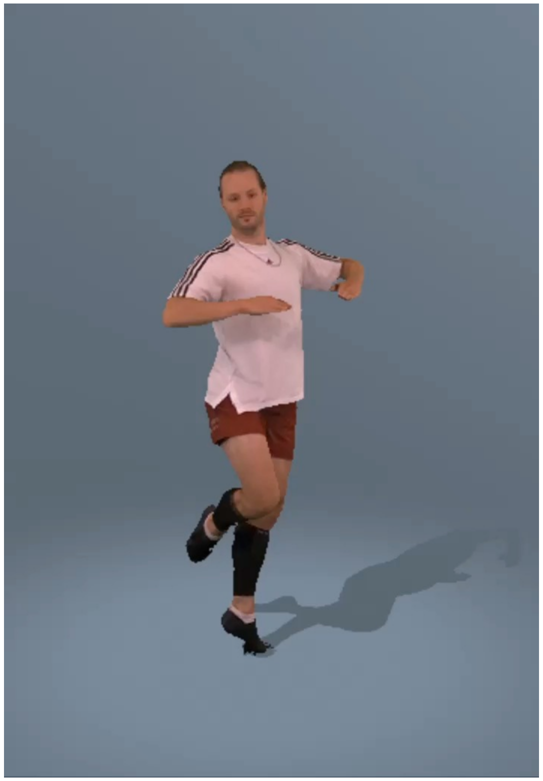
- directly predicts long-term cell visibility for Point Cloud Video (PCV).
- leverages both spatial and temporal dynamics of PCV objects and viewers.
- overcomes Limitations of Trajectory-Based Methods.
- enhances long-term 6-DoF FoV prediction and benefits immersive video streaming and 3D rendering.
- *code publicly available:* https://github.com/chenli1996/CellSight

Future Directions:

- more diverse content features
- collaborative group prediction
- algorithm optimization, esp. HPR
- integration with streaming bandwidth allocation

Preview: NYU Dancer 3D Video and FoV Datasets





Thanks & Questions?