Present ID # 435
Paper ID # 1082

Layout Sequence Prediction From Noisy Mobile Modality

Haichao Zhang[1], Yi Xu[1], Hongsheng Lu[2], Takayuki Shimizu[2], Yun Fu[1]

[1]Northeastern University [2]Toyota Motor North America

Layout Sequence Trajectory Beyond Vision

- **Obstructions and Object Visibility**: How can we predict object trajectories effectively when the camera is obstructed, and objects temporarily vanish from sight?
  Combining Vision and Mobile Computing.
- **Size Inference from Incomplete Trajectories**: Is it feasible to accurately infer missing object size information from incomplete trajectories and sensors' signals?
  Layout Sequence Trajectory Prediction.

Introduction

❖ Motivation

Real-world situations often involve obstructed cameras, missed objects, or objects that are out of sight due to environmental factors, resulting in incomplete or noisy trajectories.

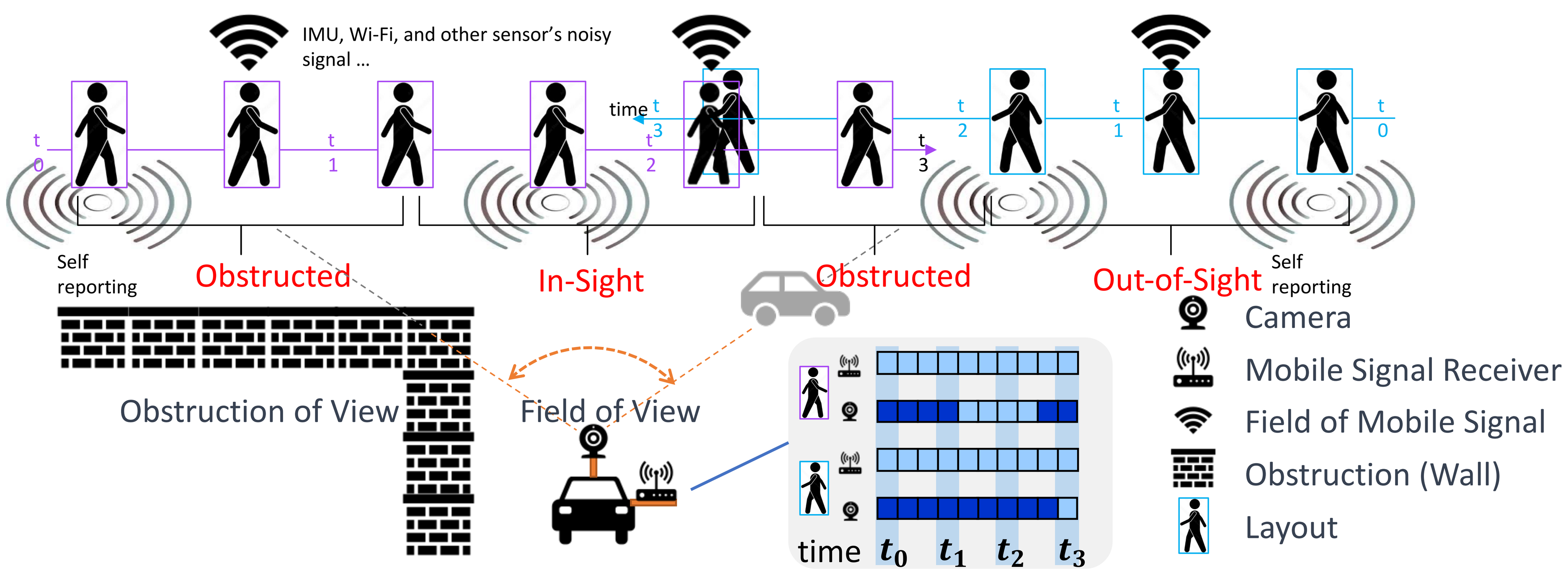

Figure 1: Real-World Scenario with Obstructed Cameras and Missing Objects

❖ Existing Methods and Drawbacks

- Computer vision, accurate but have a limited observation range and suffer from obstruction problems.
- Mobile computing doesn't suffer from out-of-sight issues but is noisy.

❖ Challenges

- Leveraging the mobile modality often introduces noise.
- Important information, such as object size and other detailed information contained in the bounding box, is often missing.

❖ Contributions

- ✓ A novel task: Combining visual and mobile modalities to enhance sequence observation range and prediction accuracy, effectively addressing their individual limitations.
- ✓ Layout sequence: Extending traditional trajectory prediction into layout sequence prediction to provide detailed object information, such as bounding boxes and depth.
- ✓ The LTrajDiff Model: Accurately predicting trajectory sequences from noisy and obstructed layout sequences, significantly improving prediction accuracy.

Our Method

❖ Overview

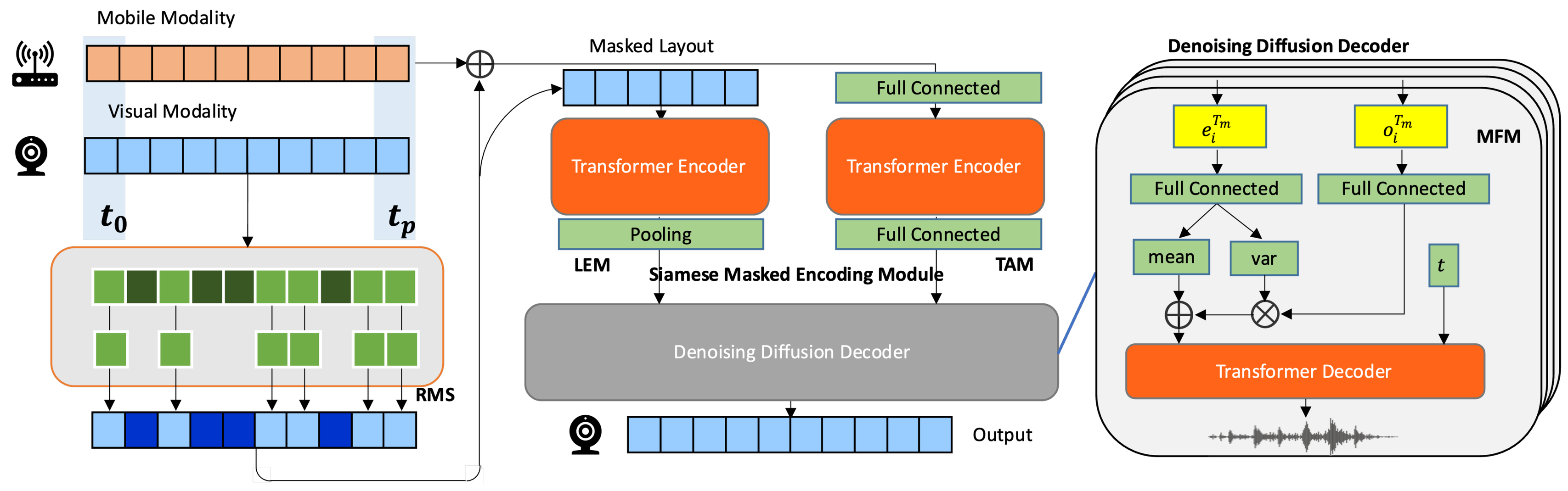

Figure 2: Overview of the proposed LTrajDiff Model

❖ Random Mask Strategy(RMS)

- Simulates masks for obstructed and out-of-sight scenarios.
- Utilizes a stochastic function ***M_i^(t_q:t_p)*** with a random variable r (sampled from $U(0,1)$) to create masks.

$$M_i^{t_q:t_p} = f_{Shuffle}\left([0]_{(q-p)*r} \circ [1]_{(q-p)*(1-r)}\right), r \sim U(0,1)$$

❖ Siamese Mask Encoding Module

Comprises two key elements:

Temporal Alignment Module(TAM)
Aligns mobile and visual modalities, extracting temporal information.

Layout Extracting Module(LEM)
Infers object size, layout, and other detailed information using unmasked layout timestamps.

❖ Denoising Diffusion Decoder

Employs a coarse-to-fine diffusion model to remove noise and generate denoised layout sequences.

Modality Fusion Module (MFM)
Jointly obtains embeddings from layout and temporal alignment features to fuse information from both modalities.

Results

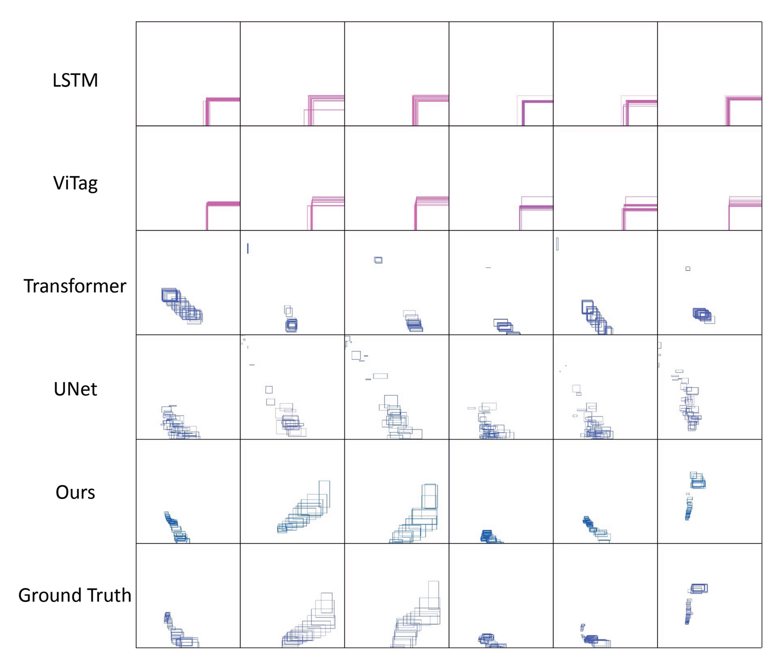

Figure 3: Visualization Results

Model	Phase I ↓	Phase II ↓
LSTM [33]	110.11	116.24
ViTag [6]	110.30	110.32
Transformer [9]	28.28	28.27
UNet [32]	3.42	5.52
HIVT [37]	-	17.51
MID [11]	-	13.32
LTrajDiff(Ours)	-	4.48

Table 1: Results on Extremely Short Inputs

Dataset	H3D [30]	Vi-Fi [18]	
Metrics	MSE-T ↓	MSE-T ↓	IoU-D ↑
LSTM [33]	452.14	432.33	0.04
ViTag [6]	455.61	421.42	0.04
Transformer [9]	5.17	58.29	0.42
UNet [32]	5.92	58.79	0.43
MID [11]	13.48	64.07	0.18
HIVT [37]	2.88	66.07	0.19
LTrajDiff (Ours)	2.72	56.13	0.69

Table 2: Results on H3D and Vi-Fi

Modality Variant	MSE-T ↓	IoU-D ↑
w/o Mobile Modality	387.86	0.29
w/o Visual Modality	362.33	0.33
Mobile + Visual Modality	56.13	0.69

Table 3: Ablation Study of Modality

Model Variant	MSE-T ↓	IoU-D ↑
w/o RMS (4.1)	307.41	0.08
w/o MFM (4.3.1)	113.55	0.45
w/o TAM (4.2.1)	295.93	0.15
w/o LEM (4.2.2)	64.07	0.18
Complete model	56.13	0.69

Table 4: Ablation Study of Model