

# ViewBirdiformer: Learning to recover ground-plane crowd trajectories and ego-motion from a single ego-centric view

Mai Nishimura<sup>1,2</sup>, Shohei Nobuhara<sup>2</sup> and Ko Nishino<sup>2</sup>

**Abstract**—We introduce a novel learning-based method for view birdification [1], the task of recovering ground-plane trajectories of pedestrians of a crowd and their observer in the same crowd just from the observed ego-centric video. View birdification becomes essential for mobile robot navigation and localization in dense crowds where the static background is hard to see and reliably track. It is challenging mainly for two reasons; i) absolute trajectories of pedestrians are entangled with the movement of the observer which needs to be decoupled from their observed relative movements in the ego-centric video, and ii) a crowd motion model describing the pedestrian movement interactions is specific to the scene yet unknown a priori. For this, we introduce a Transformer-based network referred to as ViewBirdiformer which implicitly models the crowd motion through self-attention and decomposes relative 2D movement observations onto the ground-plane trajectories of the crowd and the camera through cross-attention between views. Most important, ViewBirdiformer achieves view birdification in a single forward pass which opens the door to accurate real-time, always-on situational awareness. Extensive experimental results demonstrate that ViewBirdiformer achieves accuracy similar to or better than state-of-the-art with three orders of magnitude reduction in execution time.

## I. INTRODUCTION

We as human beings have a fairly accurate idea of the absolute movements of our surroundings in the world coordinate frame, even when we can only observe their movements relative to our own in our sight such as when walking in a crowd. Enabling a mobile agent to maintain a dynamically updated map of surrounding absolute movements on the ground, solely from observations collected from its own vantage point, would be of significant use for various applications including robot navigation [2], autonomous driving [3], sports analysis [4], and crowd monitoring [5]–[7]. The key challenge lies in the fact that when the observer (*e.g.*, person or robot) is surrounded by other dynamic agents, static “background” can hardly be found in the agent’s field of view. In such scenes, conventional visual localization methods including SLAM would fail since static landmarks become untrackable or unreliable due to frequent occlusions and observation noises sensitive to distances [8]. External odometry signals such as IMU and GPS are also often unreliable. Even when they are available, visual feedback becomes essential for robust pose estimation (imagine walking in a crowd with closed eyes).

Nishimura *et al.* recently introduced this exact task as *view birdification* whose goal is to recover on-ground

trajectories of a camera and a crowd just from perceived movements (not appearance) in an ego-centric video [1]<sup>1</sup>. They proposed to decompose these two types of trajectories, one of the pedestrians in the crowd and another of a person or mobile robot with an ego-view camera, with a cascaded optimization which alternates between estimating the displacements of the camera and estimating those of surrounding pedestrians while constraining the crowd trajectories with a pre-determined crowd motion model [9], [10]. This iterative approach suffers from two critical problems which hinder their practical use. First, its iterative optimization incurs a large computational cost which precludes real-time use. Second, the analytical crowd model as a prior is restricting and not applicable to diverse scenes where the crowd motion model is unknown.

In this paper, we propose *ViewBirdiformer*, a Transformer-based view birdification method. Instead of relying on restrictive assumptions on the motion of surrounding people and costly alternating optimization, we define a Transformer-based network that learns to reconstruct on-ground trajectories of the surrounding pedestrians and the camera from a single ego-centric video while simultaneously learning their motion models. As Fig. 1 depicts, ViewBirdiformer takes in-image 2D pedestrian movements as inputs, and outputs 2D pedestrian trajectories and the observer’s ego-motion on the ground plane. The multi-head self-attention on the motion feature embeddings of each pedestrian of ViewBirdiformer captures the local and global interactions of pedestrians. At the same time, it learns to reconstruct on-ground trajectories from observed 2D motion in the image with cross-attention on features coming from different viewpoints.

A key challenge of this data-driven view birdification lies in the inconsistency of coordinate frames between input and output movements—the input is 2D in-image movements of pedestrians relative to ego-motion, but the expected outputs are on-ground trajectories in absolute coordinates (*i.e.*, independent of the observer’s motion). ViewBirdiformer resolves this by introducing the two types of queries, *i.e.*, the camera ego-motion and pedestrian trajectories, in a multi-task learning formulation, and by transforming coordinates of pedestrian queries relative to the previous ego-motion estimates.

We thoroughly evaluate the effectiveness of our method using the view birdification dataset [1] and also by conducting ablation studies which validate its key

<sup>1</sup>Mai Nishimura is with OMRON SINIC X Corporation, 5-24-5, Hongo, Bunkyo-ku, Tokyo, Japan [mai.nishimura@sinicx.com](mailto:mai.nishimura@sinicx.com). <sup>2</sup>Shohei Nobuhara and Ko Nishino are with Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, Japan [i.kyoto-u.ac.jp](mailto:{nob, kon}@i.kyoto-u.ac.jp)

<sup>1</sup>Note that Bird’s Eye View transform is a completely different problem as it concerns a single frame view of the appearance (not the movements) and cannot reconstruct the camera ego-motion.

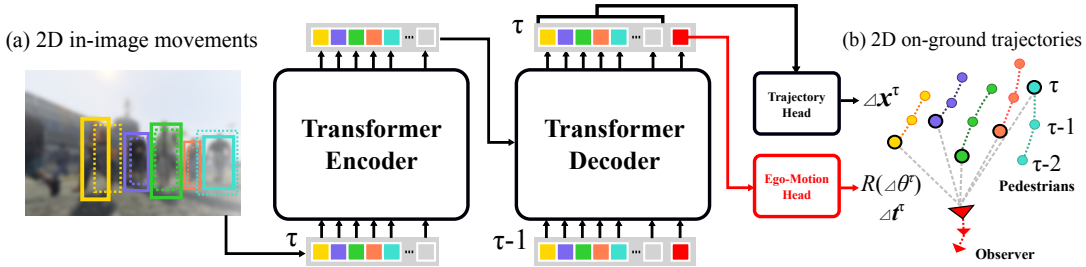


Fig. 1: Given bounding boxes of moving pedestrians in an ego-centric view captured in the crowd, ViewBirdiformer reconstructs on-ground trajectories of both the observer and the surrounding pedestrians.

components. The proposed Transformer-based architecture learns to reconstruct trajectories of the camera and the crowd while learning their motion models by adaptively attending to movement features of them in the image plane and on the ground. It enables real-time view birdification of arbitrary ego-view crowd sequences in a single inference pass, which leads to three orders of magnitude speedup from the iterative optimization approach [1]. We show that the results of ViewBirdiformer can be opportunistically refined with geometric post-processing, which results in similar or better accuracy than state-of-the-art [1] but still in orders of magnitude faster execution time.

## II. RELATED WORK

### A. View Birdification

As summarized in Table I, View Birdification [1] is not the same as bird’s-eye view (BEV) transformation [11]–[14]. BEV transformation refers to the task of rendering a 2D top-down view image from an on-ground ego-centric view and concerns the appearance of the surroundings as seen from the top and does not resolve the ego-motion, *i.e.*, all recovered BEVs are still relative to the observer. View birdification, in contrast, reconstructs both the observer’s and surrounding pedestrians’ locations on the ground so that the relative movements captured in the ego-centric view can be analyzed in a single world coordinate frame on the ground (*i.e.*, “birdified”). View birdification thus fundamentally differs from BEV transform as it is inherently a 3D transform that accounts for the ego-motion, *i.e.*, the 2D projections of surrounding people in the 2D ego-view need to be implicitly or explicitly lifted into 3D and translated to cancel out the jointly estimated ego-motion of the observer before being projected down onto the ground-plane. Nishimura *et al.* introduced a geometric method for view birdification [1], which explicitly transforms the 2D projected pedestrian movements into 3D but on the ground plane with a graph energy minimization by leveraging analytically expressible crowd motion models [9]. Our method fundamentally differs from this in that the transformation from 2D in-image movement to on-ground motion as well as the on-ground coordination of pedestrian motion is jointly learned from data.

TABLE I: View birdification (VB) is the only task that simultaneously recovers the absolute trajectories of the camera and its surrounding pedestrians only from their perceived movements relative to an observer.

Task	Input		Output		Scenes	
	static	dynamic	ego	traj	a few people	crowd
BEV [11]	✓				✓	
3D MOT [15]		✓		✓	✓	✓
SLAM [16]	✓	✓	✓	✓	✓	
VB (Ours, [1])		✓	✓	✓	✓	✓

### B. Simultaneous Localization and Mapping (SLAM)

Dynamic SLAM and its variants inherently rely on the assumption that the world is static [17], [18]. Dynamic objects cause feature points to drift and contaminate the ego-motion estimate and consequently the 3D reconstruction. Past methods have made SLAM applicable to dynamic scenes, “despite” these dynamic objects, by treating them as outliers [19] or explicitly tracking and filtering them [20]–[23]. A notable exception is Dynamic Object SLAM which explicitly incorporates such objects into its geometric optimization [16], [24], [25]. The method detects and tracks dynamic objects together with static keypoints, but assumes that the dynamic objects in view are rigid and obey a simple motion model that results in smoothly changing poses. None of the above methods consider the complex pedestrian interactions in the crowd [5], [9], [26], [27]. Our method fundamentally differs from dynamic SLAM in that it reconstructs both the observer’s ego-motion and the on-ground trajectories of surrounding dynamic objects without relying on any static key-point, while also recovering the interaction between surrounding dynamic objects. In other words, the movements themselves are the features.

### C. 3D Multi-Object Tracking (3D MOT)

3D MOT concerns the detection and tracking of target objects in a video sequence while estimating their 3D locations on the ground [28], [29]. Most recent works aim to improve tracklet association across frames [29]. These approaches, however, assume a simple motion model independent of the camera ego-motion [30], which hardly applies to a dynamic observer in a crowd with complex interactions with other pedestrians. 3D MOT in a video

with a dynamic observer [15] has been studied, but the observer motion is known from an external GPS which is often inaccurate [31]. Our work focuses on reconstructing both the observer ego-motion and surrounding pedestrians in a crowd, while simultaneously learning their complex interactions, which complements these works for visual situational awareness and surveillance.

### III. VIEW BIRDIFICATION

Let us first review the task of view birdification [1]. We have a crowd of people and one observer in the crowd with an ego-centric camera observing the surroundings while moving around. The observer can either be one of the pedestrians of the crowd or a mobile robot, or even an autonomous vehicle, in the crowd. As the observer is immersed in the crowd with a limited but dynamic field-of-view, the static background cannot be reliably found in the ego-centric view.

Let us assume that the crowd consists of  $N$  people. We set the z-axis of the world coordinate system to the normal of the ground plane (xy-plane). As in previous work [1], we assume the ground plane to be planar and the observer’s camera direction is parallel to it. We can assume this without loss of generality as the camera pitch and roll can be corrected either by measurements of the moment (*e.g.*, with an IMU) [32]. View birdification thus is the problem of recovering 2D trajectories of the observer and surrounding people (visible in the ego-centric view) on the ground plane (xy-plane) from their 2D in-image movements in the ego-centric view.

Let  $\mathbf{x}_i = [x_i, y_i]^\top$  denote the on-ground location of the  $i^{\text{th}}$  pedestrian and  $\boldsymbol{\pi} = [c_x, c_y, \theta_z]$  the pose of the observer’s camera. The ego-centric camera pose  $\boldsymbol{\pi}$  consists of a rotation matrix  $R(\theta_z) \in \mathbb{R}^{2 \times 2}$  parameterized by the rotation angle around the z-axis  $\theta_z$  and 2D translation  $\mathbf{t} = -R(\theta_z) [c_x, c_y]^\top$ , *i.e.*, the viewing direction and camera location on the ground, respectively. The observer’s camera location is  $[c_x, c_y, c_z]^\top$ , where the mounted height  $c_z$  is constant across frames, and the intrinsic matrix  $A \in \mathbb{R}^{3 \times 3}$  is assumed to be constant.

At every timestep  $\tau$ , we extract the state of each pedestrian  $\mathbf{s}_i^\tau$  for all those visible in the observed image,  $n \in \{1, 2, \dots, N\}$ . The pedestrian state encodes the two-dimensional center of the pedestrian’s bounding box and the velocity calculated by its displacement from the bounding box center of the previous  $(\tau - 1)$  frame. These states of visible pedestrians in the ego-centric view  $\mathcal{S}_i^{\tau_1:\tau_2}$  can be extracted with an off-the-shelf multi-object tracker with consistent IDs. Given a sequence of in-image pedestrian states  $\mathcal{S}_i^{\tau_1:\tau_2} = \{\mathbf{s}_i^{\tau_1}, \mathbf{s}_i^{\tau_1+1}, \dots, \mathbf{s}_i^{\tau_2}\}$  from timestep  $\tau_1$  to  $\tau_2$ , our goal is to simultaneously reconstruct the on-ground trajectories of pedestrians  $\mathcal{X}_i^{\tau_1:\tau_2} = \{\mathbf{x}_i^{\tau_1}, \mathbf{x}_i^{\tau_1+1}, \dots, \mathbf{x}_i^{\tau_2}\}$  and the observer’s camera poses  $\Pi^{\tau_1:\tau_2} = \{\boldsymbol{\pi}^{\tau_1}, \boldsymbol{\pi}^{\tau_1+1}, \dots, \boldsymbol{\pi}^{\tau_2}\}$ .

### IV. VIEWBIRDFORMER

Our goal is to devise a method that jointly transforms the 2D in-image movements into 2D on-ground trajectories and models the on-ground interactions between pedestrians

in a single framework. For this, we formulate view birdification as a set-to-set translation task, and derive a novel Transformer-based network referred to as *ViewBirdiformer*.

#### A. Geometric 2D-to-2D Transformer

Given a sequence of in-image pedestrian states for  $N$  people in a crowd at time  $\tau$ , we first embed them into a set of  $d$ -dimensional state feature vectors  $\mathcal{F}_s \in \mathbb{R}^{N \times d}$  with a multilayer perceptron (MLP). We similarly embed past  $(\tau - 1)$  on-ground trajectories of the pedestrians and the observer’s camera, too. ViewBirdiformer consists of an encoder that encodes input in-image state features  $\mathcal{F}_s$  into a sequence of hidden state features  $\mathcal{H}_s \in \mathbb{R}^{N \times d}$ , and a decoder that takes in the hidden features and on-ground queries  $\mathcal{Q}_o \in \mathbb{R}^{(N+1) \times d}$

$$\mathcal{H}_s = \mathcal{E}_\psi(\mathcal{F}_s), \quad \mathcal{F}_o = \mathcal{D}_\phi(\mathcal{Q}_o, \mathcal{H}_s), \quad (1)$$

where  $\mathcal{E}_\psi$  and  $\mathcal{D}_\phi$  are the encoder and decoder models with learnable model parameters  $\psi$  and  $\phi$ , respectively. Figure 2 depicts the overall architecture of our ViewBirdiformer.

*a) Attention layers:* A standard attention mechanism [33] accepts three types of inputs: a set of queries  $\mathcal{Q} \in \mathbb{R}^{M \times d}$ , a set of key vectors  $\mathcal{K} \in \mathbb{R}^{N \times d}$ , and a set of value embeddings  $\mathcal{V} \in \mathbb{R}^{N \times d}$ . The output is computed by values weighted by an attention matrix  $\mathbf{A} \in \mathbb{R}^{M \times N}$  composed of dot-products of queries and keys, and we use softmax to normalize the attention weights,

$$\text{Attn}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \sum_{j=1}^N A_{ij} \mathbf{v}_j, \quad \mathbf{A}_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{j'=1}^N \exp(\mathbf{q}_i^\top \mathbf{k}_{j'})} \quad (2)$$

The query  $\mathbf{q}$ , key  $\mathbf{k}$ , and value  $\mathbf{v}$  vectors are linear embeddings of the source  $\mathbf{f}_s$  and target  $\mathbf{f}_t$  input state features

$$\mathbf{q} = W_q(\mathbf{f}_t), \quad \mathbf{k} = W_k(\mathbf{f}_s), \quad \mathbf{v} = W_v(\mathbf{f}_s), \quad (3)$$

where  $W_q$ ,  $W_k$ , and  $W_v$  are linear embedding matrices specific to the vector types. We refer to the case of  $\mathbf{f}_s = \mathbf{f}_t$  as *self-attention*, and the other case  $\mathbf{f}_s \neq \mathbf{f}_t$  as *cross-attention*.

*b) Attention Mask:* To handle the varying number of pedestrians entering and leaving the observer’s view, we apply a mask  $\mathbf{M}^\tau \in \mathbb{R}^{M \times N}$  to the attention matrix  $\mathbf{A}^\tau$  as  $\mathbf{M}^\tau \odot \mathbf{A}^\tau$ , where  $\odot$  denotes Hadamard product. The element of the mask  $M_{ij}^\tau$  is set to  $\mathbf{0}$  if either of the pedestrians  $i$  or  $j$  are missing at time  $\tau$ , otherwise  $\mathbf{1}$ . This allows us to handle temporarily occluded pedestrians.

*c) In-Image Motion Encoder:* The encoder architecture consists of a single multi-head self-attention layer [33] and a feed-forward network (FFN) layer. We define the input pedestrian states as  $\mathbf{s}_i^\tau = [p_x, p_y, \Delta p_x, \Delta p_y]^\top$ , consisting of the 2D center of the detected bounding box  $\mathbf{p} = [p_x, p_y]^\top$  and its velocity  $\Delta \mathbf{p} = [\Delta p_x, \Delta p_y]^\top$ . The encoder  $\mathcal{E}_\psi$  computes self-attention over all queries generated by input state feature embeddings  $\mathcal{F}_s$ , which encodes the interactions between observed pedestrians in image space.

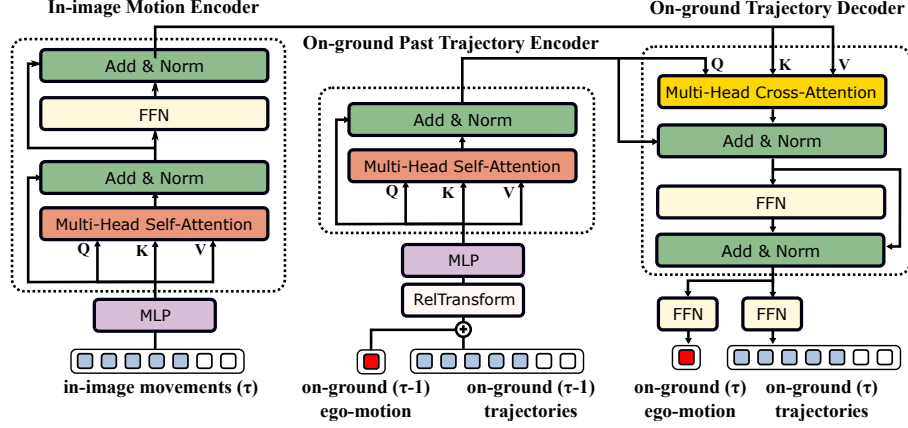


Fig. 2: **The overall architecture of ViewBirdiformer.** The decoder takes two types of queries: camera queries and pedestrian queries. These queries are fed autoregressively from the previous frame output embeddings of the last decoding layer.

d) *On-ground Trajectory Decoder:* The Transformer decoder  $\mathcal{D}_\phi$  integrates the self-attention based on-ground motion model and the cross-attention between on-ground and ego-views. First, the **On-Ground Past Trajectory Encoder** applies self-attention over queries  $\mathcal{Q}$  consisting of an ego-motion query  $q_\pi^{\tau-1}$  and on-ground pedestrian trajectory queries  $\{q_1^{\tau-1}, \dots, q_N^{\tau-1}\}$  extracted from previous estimates at  $\tau-1$ . We calculate these with on-ground queries  $q_\pi^{\tau-1} = W_q(\text{MLP}(\Delta\pi^{\tau-1}))$  and  $q_n^{\tau-1} = W_q(\text{MLP}(\mathbf{x}_i^{\tau-1} \oplus \Delta\mathbf{x}_i^{\tau-1}))$ , respectively. The attention learns to capture the implicit local and global interactions of all pedestrians to better predict the future location from past trajectories. Second, the cross-attention layer accepts hidden state features  $\mathcal{H}_s$  processed by the encoder and on-ground trajectory queries  $\mathcal{Q}_o$  processed by the self-attention layer. This layer outputs feature embeddings  $\mathcal{F}_o \in \mathbb{R}^{(N+1) \times d}$  by incorporating features from the ego-centric view. The output  $\mathcal{F}_o$  is decoded to the camera ego-motion  $\Pi^\tau$  and  $N$  pedestrian trajectories  $\{\mathcal{X}_1^\tau, \dots, \mathcal{X}_N^\tau\}$  by task-specific heads. The trajectory decoder is autoregressive, which outputs trajectory estimates one step at a time and feeds the current estimates back into the model as queries to produce the trajectories of the next timestep.

### B. Relative Position Transformation

A key challenge of view birdification lies in the inconsistency of coordinate systems between input and output trajectories. Unlike conventional frame-by-frame 2D-to-3D lifting [34] or image-based bird's eye-view transformation [11], once the viewpoint of the observer's camera is changed, the observed movements of pedestrians in the image change dramatically. To encourage the network to generalize over diverse combinations of trajectories and observer positions, we transform all the on-ground pedestrian queries relative to the previous  $\tau-1$  observer's camera estimates at every timestep,

$$\tilde{\mathbf{x}}_i^\tau = R(\theta_z^{\tau-1})\mathbf{x}_i^\tau + \mathbf{t}^{\tau-1}, \quad (4)$$

$$\Delta\tilde{\mathbf{x}}_i^\tau = R(\theta_z^{\tau-1})(\mathbf{x}_i^\tau - \mathbf{x}_i^{\tau-1}), \quad (5)$$

where  $\mathbf{t}^{\tau-1} = -R(\theta_z^{\tau-1})[c_x^{\tau-1}, c_y^{\tau-1}]^\top$  is the camera translation. We force all on-ground trajectory coordinates to be

centered on the observer's camera by defining positions and velocities relative to the observer's camera  $\tilde{\mathbf{x}} \oplus \Delta\tilde{\mathbf{x}}$  as pedestrian features, and the camera displacements  $\Delta\pi = [\Delta c_x, \Delta c_y, \Delta\theta_z]^\top$  as the observer's feature.

### C. Ego-motion Estimation by Task-specific Heads

To achieve simultaneous recovery of pedestrian trajectories and ego-motion of the observer's camera, we formulate birdification as a multi-task learning problem. Given a set of past queries  $\{q_c^{\tau-1}, q_1^{\tau-1}, \dots, q_N^{\tau-1}\}$  consisting of trajectories of the observer and surrounding pedestrians, the decoder transforms the joint set of camera and pedestrian queries into output embeddings  $\mathcal{F}_o \in \mathbb{R}^{(N+1) \times d}$ . The output embeddings  $\mathcal{F}_o$  consist of two types of features: (i) ego-motion embedding  $\mathcal{F}_{\text{ego}} \in \mathbb{R}^{1 \times d}$  from which the motion of the observer's camera on the ground is recovered, and (ii) pedestrian trajectory embeddings  $\mathcal{F}_{\text{traj}} \in \mathbb{R}^{N \times d}$  represented in a relative coordinate system, where the origin is the position of the camera. These two queries calculated from the previous  $t-1$  frame are decoded simultaneously. We define individual loss functions for these two tasks.

a) *Ego-Motion Loss:* The ego-motion output embedding  $\mathcal{F}_{\text{ego}}$  is decoded into  $\Delta\pi = [\Delta c_x, \Delta c_y, \Delta\theta_z]^\top$  by a single feed-forward network. For a batch  $\{\Delta\pi^\tau, \dots, \Delta\pi^T\} \in \mathbb{R}^{3 \times T}$  of duration  $T$ , we compute the mean squared error

$$\mathcal{L}_{\text{ego}} = \sum_{\tau=1}^T \|\Delta\hat{\pi}^\tau - \Delta\pi^\tau\|, \quad (6)$$

where  $\hat{\pi}$  is the ground-truth camera pose of an observer.

b) *Pedestrian Trajectory Loss:* Pedestrian trajectory embeddings  $\mathcal{F}_{\text{traj}}$  are decoded into 2D positions and velocities  $\tilde{\mathbf{x}} \oplus \Delta\tilde{\mathbf{x}} \in \mathbb{R}^4$  relative to the observer's camera. Given a batch of  $N$  observed pedestrians for duration  $T$ , we define the trajectory loss function as

$$\mathcal{L}_{\text{traj}} = \sum_{\tau=1}^T \sum_{i=1}^N \|\hat{\mathbf{x}}_i^\tau - (R(\theta_z^{\tau-1})^\top (\tilde{\mathbf{x}}_i^\tau - \mathbf{t}^{\tau-1}))\| + \|\Delta\hat{\mathbf{x}}_i^\tau - R(\theta_z^{\tau-1})^\top \Delta\tilde{\mathbf{x}}_i^\tau\|, \quad (7)$$

where the output estimate  $\mathbf{x}_i^\tau$  is transformed into the world coordinate system by the camera pose estimates consisting of the rotation angle  $\theta_z^\tau = \theta_z^{\tau-1} + \Delta\theta_z^\tau$  and 2D translation  $\mathbf{t}^\tau = R(\Delta\theta_z^\tau)\mathbf{t}^{\tau-1} + \Delta\mathbf{t}^\tau$ .

c) *Observer Reprojection Loss*: What makes view birdification unique from other on-ground trajectory modeling problems is its ego-centric view input. Although the 2D ego-centric view degenerates depth information of the observed pedestrian movements, it also provides a powerful inductive bias for on-ground trajectory estimates. Using the observer’s camera intrinsic matrix  $A$ , we compute the reprojection loss in the image plane

$$\mathcal{L}_{\text{proj}} = \sum_{\tau=1}^T \sum_{i=1}^N \|\bar{\mathbf{p}}_i^\tau - sA\bar{\mathbf{x}}_i^\tau\|, \quad (8)$$

where  $\bar{\mathbf{p}} = [p_x, p_y, 1]^\top$  is the homogeneous coordinate of the observed 2D bounding box center, and  $\bar{\mathbf{x}} = [x, y, h/2]^\top$  is the half point of the pedestrian height standing on the position  $\mathbf{x}_i = R(\Delta\theta_z)\bar{\mathbf{x}}_i + \Delta\mathbf{t}$ , respectively. The scaling factor  $s$  is determined by normalizing the  $z$ -value of the projected point in the image.

d) *Total Loss*: The complete multi-task loss becomes

$$\mathcal{L} = \mathcal{L}_{\text{tra}} + \lambda_1 \mathcal{L}_{\text{ego}} + \lambda_2 \mathcal{L}_{\text{proj}}. \quad (9)$$

To facilitate stable training, we apply curriculum learning to the reprojection loss weight  $\lambda_2$ . We set  $\lambda_2 = 0$  for the first 200 epochs, and switch to  $\lambda_2 > 0$  for the rest of the epochs.

e) *Test-time refinement*: The reprojection loss can be used to refine the ego-motion towards the pedestrian trajectory estimates at inference time. That is, we incorporate the reprojection errors into our network as a soft geometric constraint *i.e.*, weighted reprojection loss, in the training phase, and as a hard geometric constraint at inference time.

## V. EXPERIMENTS

### A. View Birdification Datasets

We evaluate our method on view birdification data consisting of paired real pedestrian trajectories and synthetic ego-views of them. The dataset is generated from public pedestrian trajectory datasets ETH [35] and UCY [36] by following the instructions of the original view birdification paper [1]. To generate a sufficient amount of ego-views including diverse patterns of projected movements, we mount two virtual, perspective cameras in front and rear on each of the pedestrians (*i.e.*, an observer) in turn. Negative heights of bounding boxes indicate the observation from a rear camera. As a result, we obtain paired trajectories and their ego-views for as many as the number of pedestrians in each scene. Following previous work [1], we assume ideal observation, *i.e.*, pedestrians are not occluded by each other and projected heights can be deduced from the observed images. There are three datasets named after the scenes they capture, **Hotel**, **ETH**, and **Students**, which correspond to sparse, moderate, and dense crowds, respectively. We prepare two types of splits of the view birdification dataset. The first one is (i) intra-scene validation split. For each scene, train,

val, and test splits are generated. This allows evaluation of how ViewBirdiformer generalizes to unseen trajectories. The second one is (ii) cross-scene validation split. We pick one scene for testing and choose the rest of the remaining scenes for validation and training. These splits allow evaluation of how ViewBirdiformer generalizes to unknown scenes.

a) *Evaluation Metric*: Our proposed framework first reconstructs the ego-motion of the observer and the trajectories of her surrounding pedestrians in the observer’s camera coordinate system. The absolute positions and trajectories of the pedestrians in the world coordinate system are computed by coupling these two outputs, *i.e.*,  $\mathbf{x}_\pi^\tau = R(\theta_z^{\tau-1} + \Delta\theta_z^\tau)\tilde{\mathbf{x}}_\pi^\tau + R(\Delta\theta_z^\tau)\mathbf{t}^{\tau-1} + \Delta\mathbf{t}^\tau$ . We evaluate the accuracy of our method by measuring the differences of the estimated positions of pedestrians  $\mathbf{x}$  and the ego-motion of the observer  $\Delta\Pi = (\Delta\mathbf{t}, \Delta\theta_z)$  from their corresponding ground truths  $\hat{\mathbf{x}}$ ,  $\hat{\Delta\mathbf{t}}$ , and  $\hat{\Delta\theta_z}$ . The translation error of the observer is  $\Delta\mathbf{t} = \frac{1}{T} \sum \|\mathbf{x}_\pi^\tau - \hat{\mathbf{x}}_\pi^\tau\|$ , where  $T$  denotes the duration of a sequence. The rotation error of the observer is  $\Delta\mathbf{r} = \frac{1}{T} \sum_\tau \arccos\left(\frac{\text{tr}(R(\Delta\theta_z^\tau)R(\hat{\Delta\theta_z^\tau})^\top) - 1}{2}\right)$ , where  $\text{tr}$  is the matrix trace. We also evaluate the absolute and relative reconstruction errors of pedestrians by  $\Delta\mathbf{x} = \frac{1}{N} \frac{1}{T} \sum_i \sum_\tau \|\mathbf{x}_i^\tau - \hat{\mathbf{x}}_i^\tau\|$  and  $\Delta\tilde{\mathbf{x}} = \frac{1}{N} \frac{1}{T} \sum_i \sum_\tau \|\tilde{\mathbf{x}}_i^\tau - R(\theta_z^{\tau-1})\hat{\mathbf{x}}_i - \mathbf{t}^{\tau-1}\|$ .

b) *Baseline Methods*: We compare our method with a purely geometric view birdification approach [1], the only other view birdification method. We use the parameter values from the original paper, which we refer to as *GeoVB-CV* and *GeoVB-SF* based on the assumed motion model: Constant Velocity (CV) [10] and Social Force (SF) [9], respectively. We also evaluate the effectiveness of the ego-view encoder and the cross-attention by comparing with the direct use of an on-ground motion model which takes  $\tau - 1$  on-ground trajectories as inputs and simply predicts positions and velocities  $\mathbf{x}^\tau \oplus \Delta\mathbf{x}^\tau$  for  $\tau$ . For this, we train a simple Transformer-based motion model with one multi-head self-attention layer which we refer to as *TransMotion*. Note that, although *ViewBirdiformer* and *GeoVB* both take as inputs the ego-centric view at time  $\tau$  and the past on-ground trajectory estimates at time  $\tau - 1$ , *TransMotion* only takes past on-ground trajectory estimates.

We consider two variants of ViewBirdiformer. The first, *ViewBirdiformer-I*, is trained on the intra-scene validation split, and the second, *ViewBirdiformer-C*, on the cross-scene validation split. Similarly, simple motion models composed of single-layer self-attention Transformers each trained with these validation splits are referred to as *TransMotion-I* and *TransMotion-C*, respectively.

c) *Implementation Details*: All networks were implemented in PyTorch. The camera intrinsic matrix  $A$  was set to that of a generic camera with FOV=120° and  $f = 2.46$ . Both the embedded dimension of the on-ground trajectories and in-image movements,  $d$  is set to 32. We use an MLP with 16 hidden units for embedding input features. The number of heads for the multi-head attention layer is all set to 8. Loss coefficient  $\lambda_1$  is set to 1.0 and  $\lambda_2$  is set to 0.3 after 200 epochs. We use Adam optimizer and set the constant

	Hotel / sparse		ETH / mid		Students / dense	
	$\Delta\tilde{x}$ [m]	$\Delta x$ [m]	$\Delta\tilde{x}$ [m]	$\Delta x$ [m]	$\Delta\tilde{x}$ [m]	$\Delta x$ [m]
TransMotion-I	-	0.183	-	0.201	-	0.216
TransMotion-C	-	0.106	-	0.223	-	0.211
GeoVB-CV [1]	<b>0.051</b>	0.070	0.089	0.115	0.023	0.024
GeoVB-SF [1]	<b>0.048*</b>	<b>0.052*</b>	<b>0.070*</b>	<b>0.079*</b>	<b>0.009*</b>	<b>0.010*</b>
<b>ViewBirdiformer-I</b>	0.123	0.123	0.170	0.170	0.071	0.071
<b>ViewBirdiformer-C</b>	0.097	0.098	0.216	0.217	0.058	0.059
<b>ViewBirdiformer-I</b> w/post-processing	0.062	0.081	0.087	0.102	<b>0.010</b>	<b>0.010*</b>
<b>ViewBirdiformer-C</b> w/post-processing	0.071	0.092	0.099	0.115	<b>0.010</b>	<b>0.011</b>
	$\Delta r$ [rad]	$\Delta t$ [m]	$\Delta r$ [rad]	$\Delta t$ [m]	$\Delta r$ [rad]	$\Delta t$ [m]
GeoVB-CV [1]	<b>0.015</b>	0.066	<b>0.016</b>	0.095	<b>0.001*</b>	<b>0.010</b>
GeoVB-SF [1]	<b>0.015</b>	<b>0.062</b>	<b>0.015*</b>	<b>0.089</b>	<b>0.001*</b>	<b>0.009*</b>
<b>ViewBirdiformer-I</b>	0.125	0.085	0.032	0.093	0.061	0.068
<b>ViewBirdiformer-C</b>	0.063	0.091	0.101	0.098	0.080	0.069
<b>ViewBirdiformer-I</b> w/post-processing	<b>0.014*</b>	<b>0.059*</b>	<b>0.015*</b>	<b>0.091</b>	<b>0.002</b>	<b>0.011</b>
<b>ViewBirdiformer-C</b> w/post-processing	<b>0.016</b>	<b>0.061</b>	0.021	0.098	<b>0.002</b>	<b>0.011</b>

TABLE II: **Quantitative Results.** The top table shows relative and absolute localization errors of pedestrian trajectories,  $\Delta\tilde{x}$  and  $\Delta x$ . The motion model baseline only extrapolates the on-ground movement and thus results in missing entries (-) in  $\Delta\tilde{x}$ . The bottom table shows the camera ego-motion errors  $\Delta r$  and  $\Delta t$ . We highlight the best (\*) and similar to best (accuracy gap  $\leq 0.005$ ) results of localization accuracy. The results demonstrate the effectiveness of our proposed ViewBirdiformer.

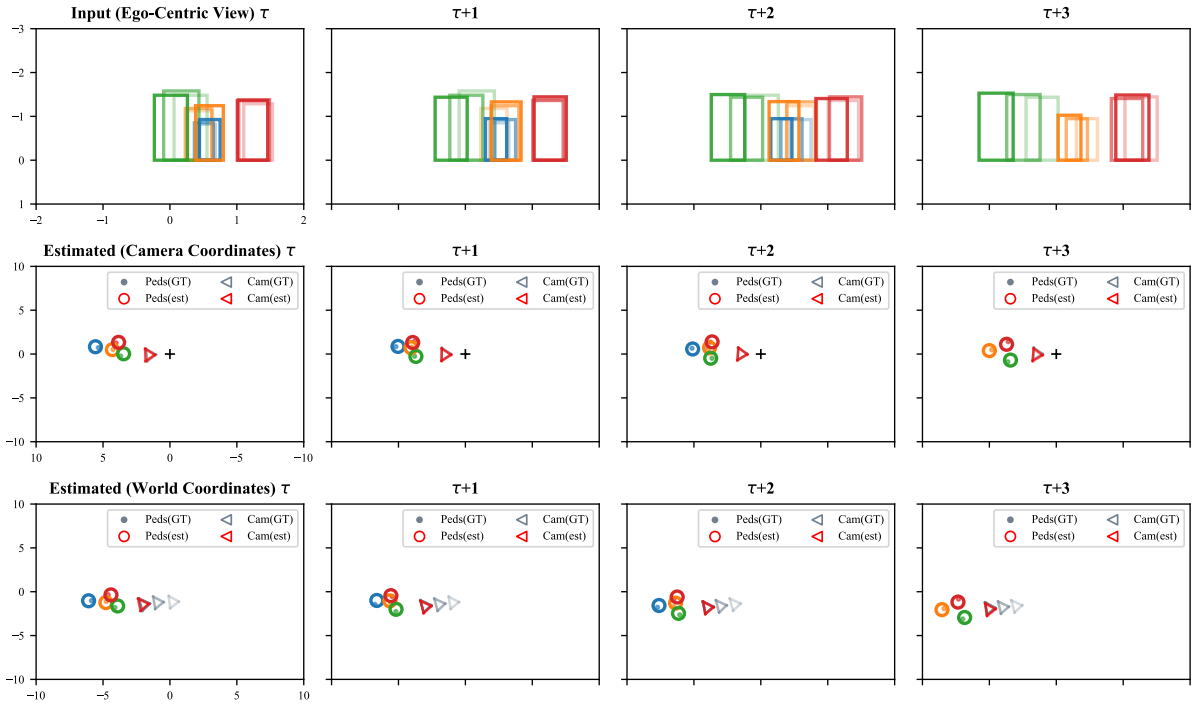


Fig. 3: **Qualitative Results of ViewBirdiformer-I without post-processing applied to ETH datasets.** The top row shows the input bounding boxes, where the same color box corresponds to the same pedestrian ID and the boxes with low alpha values correspond to the past  $\tau - 1$  frame positions. The second row shows the reconstructed camera pose and pedestrian locations at time  $\tau$  in the  $\tau - 1$  camera-centric coordinates. “+” depicts the origin of the camera coordinate system. These relative observations are converted to the world coordinates by the estimated camera pose at every frame (the third row). Grey triangles and circles denote ground-truth camera and pedestrian positions, respectively.

Dataset	Hotel / sparse			ETH / mid			Students / dense		
	$\Delta\tilde{x}$ [m]	$\Delta r$ [rad]	$\Delta t$ [m]	$\Delta\tilde{x}$ [m]	$\Delta r$ [rad]	$\Delta t$ [m]	$\Delta\tilde{x}$ [m]	$\Delta r$ [rad]	$\Delta t$ [m]
w/o RelTransform	2.115	0.055	0.277	2.105	0.053	0.279	1.713	0.090	0.269
w/o ReprojectionLoss	0.148	0.148	0.197	0.180	0.038	0.179	0.081	0.065	0.111
<b>Ours</b>	<b>0.123</b>	<b>0.125</b>	<b>0.085</b>	<b>0.170</b>	<b>0.032</b>	<b>0.093</b>	<b>0.071</b>	<b>0.061</b>	<b>0.068</b>

TABLE III: **Ablation Studies.** w/o denotes our proposed architecture without the specified component. The results demonstrate that relative transformation of the decoder inputs (Section IV-B) is essential for accurate localization of surrounding pedestrians, and the additional reprojection loss is key to accurate ego-motion estimation.

learning rate to 0.001 for all epochs. All the models are trained with a single NVIDIA Tesla V100 GPU and Intel Xeon Gold 6252 CPU. The training time is approximately 3 hours for the train split excluding Students and 14 hours for that including Students. For all the datasets, we transformed trajectories into scene-centered coordinates so that the origin of the mean position of all the pedestrians is 0. The outputs of our proposed network are post-processed by the test-time refinement described in Sec. IV-C.

### B. Comparison with Geometric Baseline

a) *Localization Accuracy:* Table II shows quantitative results. GeoVB [1] achieves high accuracy by iteratively optimizing the camera ego-motion and pedestrian positions by densely sampling possible positions for every frame. Although the accuracy of our ViewBirdiformer is slightly lower, it achieves sufficiently high absolute accuracy (e.g., 5cm errors in  $20 \times 20$  m field) with a single inference pass. Figure 3 visualizes qualitative results of our method on a typical crowd sequence, which clearly shows that our method reconstructs accurate on-ground trajectories. Even with the cross-scene validation split, *ViewBirdiformer-C* achieves comparable results. By incorporating the geometric refinement at inference time, ViewBirdiformer achieves comparable or superior accuracy to the state-of-the-art [1] but still in three orders of magnitude shorter time.

b) *Execution Time:* Figure 4 shows the execution time of our method and GeoVB [1] on a single Intel Core i5-7500 CPU and a NVIDIA GeForce 1080Ti GPU. These results clearly demonstrate the efficiency of our method compared to GeoVB. The unified transformer architecture of our method enables estimation of both ego-motion and pedestrian trajectories with a single inference pass without the costly iteration process in GeoVB [1], which results in three orders of magnitude improvement in execution time. For  $N$  pedestrians,  $S$  samples, and  $T$  iterations, the computational complexity of GeoVB is  $\mathcal{O}(NS^2T)$  and it is hardly parallelizable as it requires sequential update over all possible samples  $S$  ( $S \gg N$ ). In contrast, the computational complexity of ViewBirdiformer is  $\mathcal{O}(N^2d)$  [33] and its implementation can naturally be parallelized within GPU, which collectively realize this significant reduction in execution time. Most important, even with the geometric refinement at inference time, ViewBirdiformer achieves accuracy on par with the state-of-the-art [1] while maintaining this orders of magnitude faster execution.

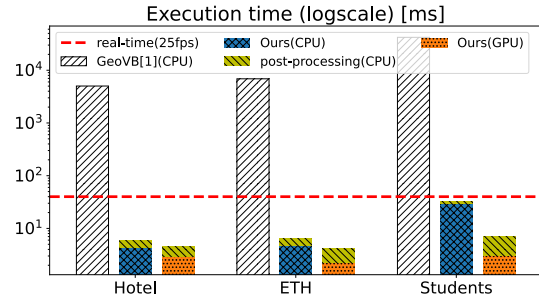


Fig. 4: **Execution time.** We measure the execution times of our method on a CPU and a GPU. The post-processing is executed on the CPU. These results are averaged over the samples.

### C. Ablation Studies

a) *Cross-Attention Between Views:* Table II compares the accuracy of ViewBirdiformer and simple extrapolation of on-ground movements using dedicated simple transformers. While ViewBirdiformer takes the current ego-centric view and the past on-ground trajectory estimates as inputs, TransMotion only takes the past trajectory estimates as inputs. ViewBirdiformer shows superior performance over TransMotion in pedestrian localization. These results clearly show that the cross-attention mechanism between on-ground motions and movements in the ego-centric views is essential for accurate trajectory estimation of the surrounding pedestrians.

b) *Relative Position Transformation:* Table III shows the results of ablating the relative position transforms (Section IV-B). All models are trained with the intra-scene split of the birdification dataset to avoid generalization errors of the learnt motion model. *w/o RelTransform* takes on-ground trajectories in world coordinates as decoder inputs. Without the relative position transformations described in Sec. IV-A, the proposed framework shows significant accuracy drops, especially in pedestrian localization. This is likely caused by the inconsistency of the coordinate system between on-ground past trajectory inputs and egocentric view inputs and demonstrates the importance of the relative transformation for generalization of the model.

c) *Reprojection Loss:* *w/o ReprojectionLoss* in Tab. III considers only the ego-motion loss and the pedestrian trajectory loss, i.e.,  $\lambda_2 = 0$  in Eq. (9). The results show that the reprojection loss slightly improves the accuracy of ego-

motion estimates. This is because the reprojection loss works similarly to geometric constraints as in *GeoVB*.

#### D. Application to Real Image Data

Supplementary video includes some results of our pre-trained model applied to a real video sequence from a MOT dataset [37]. Our method can synthetically generate data with arbitrary FOV from the crowd dataset by using only bounding box representations. This allows us to re-train models for cameras of diverse settings quite easily.

#### E. Limitations and Degenerate Scenario

If the observed relative movements are static (*i.e.*, an observer is following the pedestrian at the same speed), our model cannot break the fundamental ambiguity. Such degenerate scenarios, however, rarely happen in crowds as there will be other pedestrians. Our method also assumes that the heights of pedestrians are more or less the same and that the detected bounding boxes are correct. We plan to relax these requirements by developing an end-to-end framework that handles both tracking and birdification while estimating uncertainty in depth arising from variances of pedestrians.

### VI. CONCLUSION

In this paper, we introduced ViewBirdiformer for view birdification. The proposed architecture enables efficient and accurate view birdification by adaptively attending to movement features of the observer and pedestrians in the image plane and on the ground. Extensive evaluations demonstrate the effectiveness of ViewBirdiformer for crowds with diverse pedestrian interactions. We believe ViewBirdiformer finds use in various applications of crowd modeling and synthesis across a wide range of disciplines.

### REFERENCES

- [1] M. Nishimura, S. Nobuhara, and K. Nishino, "View birdification in the crowd: Ground-plane localization from perceived movements," in *Proc. BMVC*, 2021.
- [2] M. Nishimura and R. Yonetani, "L2b: Learning to balance the safety-efficiency trade-off in interactive crowd-aware robot navigation," in *Proc. IROS*, 2020, pp. 11 004–11 010.
- [3] K.-H. Lee, K. Matthew, G. Adrien, L. Jie, F. Chao, P. Sudeep, and B. Wolfram, "Pillarflow: End-to-end birds-eye-view flow estimation for autonomous driving," in *Proc. IROS*, 2020.
- [4] A. Cioppa, A. Deliege, F. Magera, S. Giancola, O. Barnich, B. Ghanem, and M. Van Droogenbroeck, "Camera calibration and player localization in soccer-net-v2 and investigation of their representations for action spotting," in *Proc. CVPR*, 2021, pp. 4537–4546.
- [5] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proc. CVPR*, 2018, pp. 2255–2264.
- [6] B. Ivanovic and M. Pavone, "The trajectory: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. ICCV*, 2019, pp. 2375–2384.
- [7] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. CVPR*. IEEE, 2009, pp. 935–942.
- [8] M. Nishimura, S. Nobuhara, and K. Nishino, "View birdification in the crowd: Ground-plane localization from perceived movements," *CoRR*, vol. abs/2111.05060, 2021. [Online]. Available: <https://arxiv.org/abs/2111.05060v2>
- [9] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.
- [10] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1696–1703, 2020.
- [11] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proc. CVPR*, 2021, pp. 15 536–15 545.
- [12] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "FIERY: Future instance segmentation in bird's-eye view from surround monocular cameras," in *Proc. ICCV*, 2021.
- [13] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proc. CVPR*, 2022, pp. 13 760–13 769.
- [14] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *Proc. ICRA*. IEEE, 2022, pp. 9200–9206.
- [15] H.-N. Hu, Y.-H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3d object tracking," *TPAMI*, 2022.
- [16] J. Huang, S. Yang, T.-J. Mu, and S.-M. Hu, "Clustervto: Clustering moving instances and estimating visual odometry for self and surroundings," in *Proc. CVPR*, 2020, pp. 2168–2177.
- [17] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [18] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Proc. ECCV*. Springer, 2014, pp. 834–849.
- [19] D. Hahnel, R. Triebel, W. Burgard, and S. Thrun, "Map building with mobile robots in dynamic environments," in *Proc. ICRA*, vol. 2. IEEE, 2003, pp. 1557–1563.
- [20] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4076–4083, 2018.
- [21] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Dslam: A semantic visual slam towards dynamic environments," in *Proc. IROS*. IEEE, 2018, pp. 1168–1174.
- [22] J. Vincent, M. Labbé, J.-S. Lauzon, F. Grondin, P.-M. Comtois-Rivet, and F. Michaud, "Dynamic object tracking and masking for visual slam," in *Proc. IROS*. IEEE, 2020, pp. 4974–4979.
- [23] I. Ballester, A. Fontán, J. Civera, K. H. Strobl, and R. Triebel, "Dot: Dynamic object tracking for visual slam," in *Proc. ICRA*, 2021, pp. 11 705–11 711.
- [24] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.
- [25] M. Henein, J. Zhang, R. Mahony, and V. Ila, "Dynamic slam: The need for speed," in *Proc. ICRA*. IEEE, 2020, pp. 2123–2129.
- [26] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics research*. Springer, 2011, pp. 3–19.
- [27] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proc. ICCV*, 2021.
- [28] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," in *Proc. ICRA*. IEEE, 2018, pp. 3508–3515.
- [29] J. Luiten, T. Fischer, and B. Leibe, "Track to reconstruct and reconstruct to track," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1803–1810, 2020.
- [30] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *Proc. IROS*. IEEE, 2020, pp. 10 359–10 366.
- [31] E. Héry, P. Xu, and P. Bonnifait, "Distributed asynchronous cooperative localization with inaccurate gnss positions," in *Proc. ITSC*, 2019, pp. 1857–1863.
- [32] S. Bell, A. Troccoli, and K. Pulli, "A non-linear filter for gyroscope-based video stabilization," in *Proc. ECCV*. Springer, 2014, pp. 294–308.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
- [34] L. Bertoni, S. Kreiss, and A. Alahi, "Monoloco: Monocular 3d pedestrian localization and uncertainty estimation," in *Proc. ICCV*, 2019, pp. 6861–6871.
- [35] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. ICCV*, 2009, pp. 261–268.
- [36] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer graphics forum*, vol. 26, no. 3, pp. 655–664, 2007.
- [37] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.