APPENDIX

A RELATED WORK

A.1 IO flow tensor and OD matrix

There are two major formats for crowd flow data in the literature. Earlier studies employed in-out (IO) flow tensor, the incoming and outgoing flows at the segments of a geographical region [14, 27]. Recent studies have taken on the origin-destination (OD) matrix, predicting flows between each pair of an origin and a destination segments, also referred to as OD forecasting [5, 18, 24]. The OD matrix maintains finer details of crowd flow, and the IO flow tensor can be computed by aggregating the flows of the OD matrix. Their definitions are provided in Appendix B.

A.2 Deep Crowd Flow Models

ST-ResNet is a widely used baseline for crowd flow prediction with IO flow tensor [27]. It captures spatial dependencies using 2D convolutions and takes temporal dependencies by sampling the historical IO flow at varying time intervals.

More studies have explored the stacked deep-model architecture, primarily using a convolutional neural network and LSTM to capture the two dependencies [14]. Similar approaches have been proposed for map-like input and output tensors, such as traffic prediction and taxi demand prediction [25, 26]. Other modules such as Gated Attention Network [19], have been explored also.

GEML [24] first introduced the OD matrix, to predict taxi demand and flow. It used a stacked CNN and LSTM to learn spatial features and capture temporal dependencies. MRSTN [17] also employed a stacked CNN and LSTM for predicting mobility on a rail system.

CrowdNet has achieved state-of-the-art performance in OD matrix prediction and IO flow prediction for bike share datasets [5]. It captures the spatial dependencies by graph convolution, and 1D convolution is subsequently applied in the time dimension to capture the temporal dependencies.

A.3 Transformers for OD Matrix Prediction

The transformers have been successfully applied to a wide range of problems in sequential, temporal, and computer vision domains [7, 21]. In video processing applications, STFormer [22] and other spatio-temporal transformers [2, 23] have been used to capture spatial and temporal dependencies in sequences of images.

For OD matrix prediction, few studies have combined transformer's self-attention and graph convolution for spatial feature extraction. HSTN [6] combines a graph convolution and multi-head attention for spacial feature extraction which is stacked onto a hybrid transformer-based module. ODformer [9] also stacks the combined spatial feature extractor on another transformer-based model for *periodicity* extraction. ODFormer implements OD attention, which uses two sets of query, key, and values to compute *origin*- and *destination*-wise self-attention. To compensate for computing two sets of attentions, it adopts a sparse attention mechanism [29].

B DEFINITIONS

Let G denote a set of tiles that covers a geographical area R, and N the number of tiles in G.

Definition B.1 (Spatial Tessellation). G is called a tessellation if the following properties hold: 1) G consists of a finite number of tiles, 2) tiles do not overlap, and 3) the union of all tiles covers R entirely. That is,

(1) $G = \{g_i : g_i \subset R\}_{i=1}^{N}$
(2) $g_i \cap g_j = \phi, \forall i \neq j$
(3) $\bigcup_{i=1}^{N} g_i = R$

An origin-destination (OD) matrix X represents the crowd flow by the cumulative mobility among the geographical tiles. Its entries are the number of individuals that moved from one tile to another during an interval. We denote the entry of X at i^{th} row and j^{th} column by $X^{(i,j)}$.

Let $\mathbf{r}_u = \left(r_1^u, r_2^u, \cdots\right)$ be the trajectory of u during an interval, where r_p^u is the p^{th} tile u visited.

Definition B.2. An OD matrix X is an $N \times N$ matrix whose entry $X^{(i,j)}$ represents the number of individuals moving from g_i to g_j during an interval, i.e.,

$$X^{(i,j)} = \sum_{p=1}^{P-1} \left| \left\{ \mathbf{r}_u : r_p^u = g_i \ \cap \ r_{p+1}^u = g_j \right\}_u \right|$$

where P is the length of the longest trajectory.

To represent crowd flow in an OD matrix format, it is only required that G is a tessellation. However, relation between the IO flow tensor and the OD matrix can be defined easily when G is a squared tessellation, i.e., a tessellation and comprised of equal-sized rectangular tiles [5]. Note that the proposed model does not require G to be a squared tessellation for handling the OD matrix prediction.

Let us assume that the square tiles form a $H \times W$ grid over R. We denote by $g^{(h,w)}$ the tile at the h^{th} row and the w^{th} column of the grid.

Definition B.3 (IO Flow Tensor). Let $Z_{in} \in \mathbb{R}^{H \times W}$ denote an in-flow matrix and $Z_{out} \in \mathbb{R}^{H \times W}$ an out-flow matrix, whose entries represent the number of individuals entering and leaving each tile, respectively, during an interval, i.e.,

$$Z_{in}^{(h,w)} = \sum_{p=1}^{P-1} \left| \{ \mathbf{r}_u : r_p^u \neq g^{(h,w)} \ \cap \ r_{p+1}^u = g^{(h,w)} \}_u \right|$$

$$Z_{out}^{(h,w)} = \sum_{p=1}^{P-1} \left| \{ \mathbf{r}_u : r_p^u = g^{(h,w)} \ \cap \ r_{p+1}^u \neq g^{(h,w)} \}_u \right|$$

The IO flow tensor $Z \in \mathbb{R}^{2 \times H \times W}$ is obtained by stacking the matrices Z_{in} and Z_{out},

Definition B.4. Each entry of Z_{in} and Z_{out} can also be considered the sum of flows that have the same destination and the same origin, respectively. Given X, Z_{in} and Z_{out} at $g^{(h,w)}$ can be written as follows:

$$Z_{in}^{(h,w)} = \sum_{i=1}^{N} X^{(i, H \times (w-1)+h)} \quad (3)$$

$$Z_{out}^{(h,w)} = \sum_{j=1}^{N} X^{(H \times (w-1)+h, j)} \quad (4)$$

Let X_t and Z_t denote the OD matrix and the IO flow tensor at a time interval t, respectively.

Definition B.5. OD matrix prediction is the task of predicting X_t given historical flows over τ intervals $\{X_{t-\tau}, \cdots, X_{t-1}\}$.

Definition B.6. In-Out Flow prediction is the task of predicting $Z_t \in \mathbb{R}^{2\times h\times w}$ given either historical OD flows $\{X_{t-\tau}, \cdots, X_{t-1}\}$ or historical IO flows $\{Z_{t-\tau}, \cdots, Z_{t-1}\}$.

Note that an OD matrix prediction model can also make predictions on the IO flow tensor based on (4).

C ALGORITHM

Algorithm 1 EODA Forward Propagation

INPUT: input and target OD flow matrix $\mathcal{X}$, Y, # of Transformer Block repetitions A, B
OUTPUT: RMSE loss
Function EODA: Efficient OD Attention
Function SA: Self-Attention
Function FW: Feed Forward
METHOD:
$\mathcal{X}^0_{spa} \leftarrow \text{Embed}(\text{Transpose}(\text{Reshape}(\mathcal{X})))$
$\mathcal{X}^0_{tem} \leftarrow \text{Embed}(\text{Reshape}(\mathcal{X}))$
for $a = 1, \ldots, A$ **do**
  $\mathcal{X}'^a_{spa} \leftarrow \text{BN}(\text{EODA}(\mathcal{X}^{a-1}_{spa}) + \mathcal{X}^{a-1}_{spa})$
  $\mathcal{X}^a_{spa} \leftarrow \text{BN}(\text{FW}(\mathcal{X}'^a_{spa}) + \mathcal{X}'^a_{spa})$
end for
for $b = 1, \ldots, B$ **do**
  $\mathcal{X}'^b_{tem} \leftarrow \text{BN}(\text{SA}(\mathcal{X}^{b-1}_{tem}) + \mathcal{X}^{b-1}_{tem})$
  $\mathcal{X}^b_{tem} \leftarrow \text{BN}(\text{FW}(\mathcal{X}'^b_{tem}) + \mathcal{X}'^b_{tem})$
end for
$\hat{\mathcal{Y}} \leftarrow \text{Reshape}(\text{Linear}(\mathcal{X}^A_{spa}) + \text{Linear}(\mathcal{X}^B_{tem}))$
$\hat{Y} \leftarrow \hat{\mathcal{Y}}_L$
return $\text{RMSE}(Y, \hat{Y})$

D OUTLINES OF BASELINE MODELS

Autoregression is one of the basic statistical approaches for modeling temporal dependencies and crowd flow prediction. ARIMA [12], for example, can produce a univariate forecast by leveraging the historical flow. For a global forecast over a region with ARIMA or other univariate models, one may assume independence among segments of the region or consider spatial dependencies between them through external feature engineering. VAR is an extension of ARIMA for multivariate forecasting. In VAR, spatial dependencies between all pairs of segments are treated equally as covariances of independent variables.

HA. a historical average of the previous m time steps.

ARIMA. a statistical model for understanding and forecasting future values in a time series.

VAR. a multivariate extension of ARIMA, which captures the pairwise relationships among all flows.

ST-ResNet. a deep convolutional model for spatio-temporal data, and the most widely used baseline in related studies.

DMVSTNet. exploits temporal, spatial, and semantic views and models correlations between regions that share similar temporal patterns [26].

SPN. infers the evolution of crowd flow by learning dynamic representations of time-varying data using an attention mechanism [14].