# Harnessing Fourier Isovists and Geodesic Interaction for Long-Term Crowd Flow Prediction

**Samuel S. Sohn**[1*] , **Seonghyeon Moon**[1] , **Honglu Zhou**[1] , **Mihee Lee**[1] , **Sejong Yoon**[2] , **Vladimir Pavlovic**[1] and **Mubbasir Kapadia**[1]

[1]Rutgers University, USA
[2]The College of New Jersey, USA
{sss286, sm2062, hz289, ml1323, vladimir, mk1353}@cs.rutgers.edu, yoons@tcnj.edu

## Abstract

With the rise in popularity of short-term Human Trajectory Prediction (HTP), Long-Term Crowd Flow Prediction (LTCFP) has been proposed to forecast crowd movement in large and complex environments. However, the input representations, models, and datasets for LTCFP are currently limited. To this end, we propose Fourier Isovists, a novel input representation based on egocentric visibility, which consistently improves all existing models. We also propose GeoInteractNet (GINet), which couples the layers between a multi-scale attention network (M-SCAN) and a convolutional encoder-decoder network (CED). M-SCAN approximates a super-resolution map of where humans are likely to interact on the way to their goals and produces multi-scale attention maps. The CED then uses these maps in either its encoder's inputs or its decoder's attention gates, which allows GINet to produce super-resolution predictions with substantially higher accuracy than existing models even with Fourier Isovists. In order to evaluate the scalability of models to large and complex environments, which the only existing LTCFP dataset is unsuitable for, a new synthetic crowd dataset with both real and synthetic environments has been generated. In its nascent state, LTCFP has much to gain from our key contributions. The Supplementary Materials, dataset, and code are available at sssohn.github.io/GeoInteractNet.

## 1 Introduction

In the present day, more than half of the world's population resides in the built environments of urban areas, and by 2050, this proportion is expected to reach as high as 68% [DESA, 2018]. Accordingly, it is now more important than ever to advance the modeling of human navigation within built environments. This would directly benefit application domains such as architectural design [Turner and Penn, 2002], transportation engineering [Sewall *et al.*, 2010], and crowd management [Bohannon, 2005], which are all crucial to maintaining people's safety on a daily basis. The failure for a built environment to do so is caused or signaled by overcrowding and can result in injuries for its occupants [Zhen *et al.*, 2008].

In order to measure the risk of overcrowding in an environment, we use *long-term crowd flow*, i.e., the time-aggregated spatial distribution of footfall during a crowd's movement [Sohn *et al.*, 2020], either computed from real trajectory data or predicted (Figure 1). Until recently, long-term crowd flow prediction (LTCFP) could only be done by first predicting the trajectories of agents in a crowd using either simulation models [Helbing and Molnar, 1995] from Computer Graphics literature or Human Trajectory Prediction (HTP) models [Alahi *et al.*, 2016] from Computer Vision literature. This general approach does not scale well to large crowds and environments, and we later show that HTP models are fundamentally unsuitable for LTCFP (Sec. 7.1). On the contrary, CAGE [Sohn *et al.*, 2020], the recent state-of-the-art (SOTA) framework, directly addresses LTCFP without considering individual agents of the crowd, making it inherently scalable to large crowds (but not necessarily to large environments).

However, CAGE faces three major challenges that hinder its generalizability. (1) Its environment representation is highly lossy when encoding environments with large size or complex geometry, such as angles and curves. (2) Its predictive model is unable to make accurate predictions for such environments. (3) The only existing LTCFP dataset is unsuitable for evaluating the scalability of models to large environments [Sohn *et al.*, 2020]. To address these challenges, we first contribute Fourier Isovists, a novel egocentric visibility representation of environments that uses additional channels to encode more spatial information at each cell. We then frame LTCFP as a super-resolution problem to exploit the extra information in Fourier Isovists and propose GeoInteractNet (GINet) to improve performance in this new framework using domain knowledge. Finally, we have generated a new synthetic crowd dataset with both real and synthetic environments to evaluate the scalability of CAGE and the proposed methodology.

## 2 Related Work

Simulation models have been developed over the past 2 decades by the Computer Graphics community. These models are typically hand-crafted to produce specific phe-

[*] Contact Author

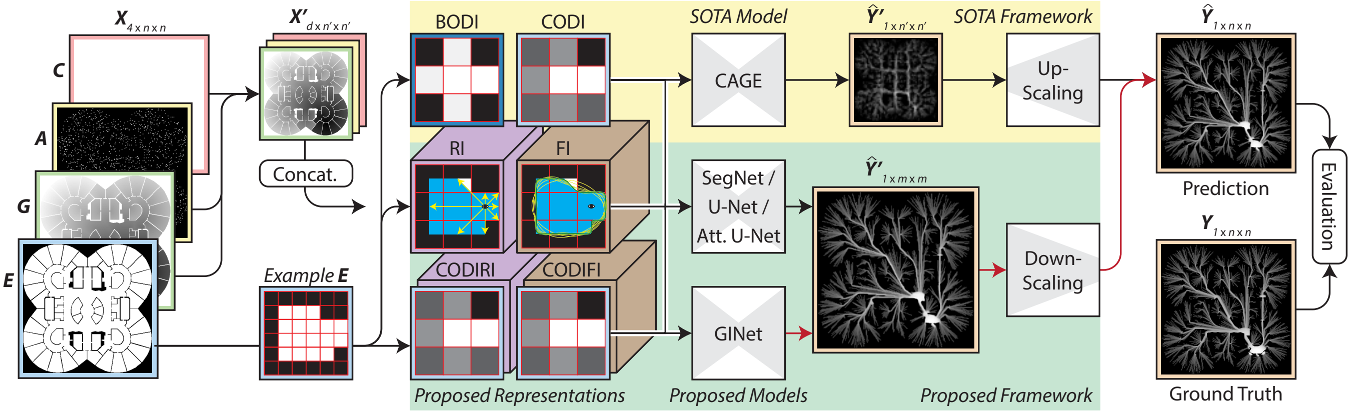

Figure 1: **Framework Overview.** Using the inputted crowd scenario, we compute 6 SOTA and novel environment representations to test with SOTA and novel frameworks. Our proposed super-resolution framework predicts a larger crowd flow map than the SOTA framework to avoid losing fine-grained details.

nomena about real crowd movement, such as social interactions [Helbing and Molnar, 1995] and collision avoidance [Van Den Berg *et al.*, 2011]. While these models are highly robust, they are only accurate with respect to real data at a high level (e.g. the rate of agent flow) since they have very few tunable parameters. In order to make long-term predictions, these models simulate each agent's movement one small step at a time. The robustness of this method has led to its adoption in application domains, but its inaccuracy and inefficiency leave much room for improvement.

On the other hand, Human Trajectory Prediction (HTP) models were introduced 5 years ago by the Computer Vision community. These models are data-driven and learn short-term trajectory distributions based on datasets of real human trajectories [Alahi *et al.*, 2016; Gupta *et al.*, 2018]. In contrast to crowd simulation models, HTP models have many learnable parameters and are more accurate with respect to real data than simulation models at a low level (e.g., short-term motion statistics) [Sohn *et al.*, 2021]. However, successive short-term predictions from these models compound error over time [Salzmann *et al.*, 2020], meaning that they cannot reliably make long-term predictions and are thereby inaccurate at a high level. These models are also unable to avoid collisions with other agents or obstacles in the environment despite conditioning their predictions on this very information [Sohn *et al.*, 2021]. While HTP models show promise, they have not yet taken the place of simulation models in application domains.

These prior works are iterative and agent-centric in nature, meaning that they scale poorly to large crowds and environments. The SOTA CAGE framework showed for the first time that accurate and efficient predictions of long-term crowd flow could be made in an instantaneous and environment-centric fashion by treating LTCFP as a vision problem and leveraging convolutional neural networks [Sohn *et al.*, 2020]. The contributions of this work are in addressing CAGE's key limitations and advancing LTCFP, which fulfills a unique role in the analysis of crowd movement that simulation and HTP models are ill-suited to conduct due to their inaccuracy and inefficiency.

## 3 Problem Definition

LTCFP is defined as the prediction of a crowd flow map $\mathbf{Y}_{1\times n\times n}$ using the initial state of a crowd and environment $\mathbf{X}_{4\times n\times n}$ (i.e., the crowd scenario), where both the input and output are represented as images with the same top-down orthogonal perspective of a $\frac{n}{s}\times\frac{n}{s}$ m$^2$ environment and $s$ is a fixed scale across all environments. The crowd flow map $\mathbf{Y}$ is computed by the log transform of a crowd's cumulative footfall while moving from its initial configuration to a shared goal position, where footfall is measured at 50 Hertz. The assumption of a shared goal makes LTCFP applicable to ingress and egress scenarios, which are of interest to application domains. The crowd scenario $\mathbf{X}=\langle\mathbf{C},\mathbf{A},\mathbf{G},\mathbf{E}\rangle$ contains the following channel information at each pixel $(i,j)$: $\mathbf{C}_{i,j}\in\{1\}$ is its amount of environment Compression (initially 1), $\mathbf{A}_{i,j}\in\{0,1\}$ is whether an Agent is occupying it, $\mathbf{G}_{i,j}\in[0,1]$ is its normalized distance to the Goal, and $\mathbf{E}_{i,j}\in\{0,1\}$ is its navigability in the Environment. In order to leverage a convolutional encoder-decoder network, the CAGE framework compresses $\mathbf{X}$, which varies in spatial size with the environment, into $\mathbf{X}'_{4\times n'\times n'}$ as input to the network, where $n'$ is fixed.[1] The network's prediction $\widehat{\mathbf{Y}}'_{1\times n'\times n'}$ is then bilinearly interpolated to $\widehat{\mathbf{Y}}_{1\times n\times n}$.

By fixing $n'$ and compressing any environment with $n>n'$, the accuracy of models is inherently limited and their predictions must be *upscaled* to $\widehat{\mathbf{Y}}$. Our proposed changes to this framework are (1) to compress the raw input into $\mathbf{X}'_{d\times n'\times n'}\mid d\geq 4$ with additional channels to store more environment information and (2) to use the extra channels for predicting a super-resolution $\widehat{\mathbf{Y}}'_{1\times m\times m}$, where $m$ is a fixed value greater than $n'$ by a factor of $2^c\mid c\geq 1$ (Figure 1). The proposed framework anticipates that $n$ may be higher than $n'$ and compensates for it by a factor of $2^c$, meaning that for any environment up to $2^c$ larger than $n'$, the resolution of the

[1] Matrices with fixed spatial size are denoted by $'$.

prediction is not a limiting factor and the prediction must be *downscaled* to $\widehat{\mathbf{Y}}$. We have set $n' = 64$ and $m = 256$ according to our dataset.

## 4 Crowd and Environment Representations

We propose several novel representations of $\mathbf{X}' = \langle \mathbf{C}', \mathbf{A}', \mathbf{G}', \mathbf{E}' \rangle$ for predicting the crowd flow $\widehat{\mathbf{Y}}'$ and compare them to the SOTA representation. All representations share the same compression, agent, and goal channels, but encode environment information $\mathbf{E}'$ differently. $\mathbf{A}'$ and $\mathbf{G}'$ are computed by bilinearly interpolating $\mathbf{A}$ and $\mathbf{G}$, and each cell of $\mathbf{C}'$ takes the value of $n'/n$. The environment representations can be divided into three categories: image, isovist, and joint image-isovist representations.

**Image Representations (BODI and CODI).** Image representations encode the environment as a single-channel image. CAGE uses a binary image representation that losslessly and non-uniformly compresses rectangular regions in environments with axis-aligned geometry [Sohn *et al.*, 2020]. However for large and complex environments it becomes equivalent to bilinearly interpolating $\mathbf{E}$ to $\mathbf{E}'$ and thresholding each pixel at 0.5, which is highly lossy. We use this Binary One-Dimensional (BODI) representation $\mathbf{E}' \in \{0, 1\}^{1 \times n' \times n'}$ as the baseline. The proposed Continuous One-Dimensional (CODI) representation $\mathbf{E}' \in [0, 1]^{1 \times n' \times n'}$ is also computed by interpolating the environment, but unlike BODI, its values are not thresholded (Figure 1).

**Isovist Representations (RI and FI).** The proposed isovist representations $\mathbf{E}' \in \mathbb{R}^{r \times n' \times n'}$ encode egocentric visibility information along a uniform grid of $n' \times n'$ viewpoints across the environment. At each viewpoint, a 360° visibility polygon (i.e., an isovist [Benedikt, 1979]) is computed and then converted into an $r$-dimensional feature vector along the channel dimension, where $r = 60$. If the viewpoint is in a non-navigable space, the isovist is represented by a zero vector. The incorporation of isovists is motivated by prior studies, which have evidenced that human navigational behavior in indoor environments is strongly correlated with isovists [Wiener *et al.*, 2007; Wiener and Franz, 2004]. Furthermore, isovists cannot normally be learned by convolutional neural networks, because they can gather visibility information from long distances that well-exceed reasonable kernel sizes, e.g., an isovist at one end of a long hallway encodes information at the opposite end.

The Raw Isovist (RI) representation computes the isovist's feature vector using a viewpoint distance function [Zhang *et al.*, 2002], which casts $r$ equiangular rays (i.e., sightlines) from the viewpoint and encodes the length of each ray when it hits the contour of the isovist. This makes RI well-suited for encoding angular geometries. On the other hand, the Fourier Isovist (FI) representation computes the isovist's feature vector as the first $\frac{r}{4}$-harmonics of elliptic Fourier features [Kuhl and Giardina, 1982]. The first harmonic represents the ellipse that best fits the contour, and each subsequent harmonic represents an ellipse that rolls around the previous harmonic's ellipse, capturing increasingly fine-grained details in a multi-scale manner. These elliptic features make FI well-suited for encoding curved geometries. Figure 1 shows incremental reconstructions of the first 15 harmonics for an example $\mathbf{E}$. Further analysis of RI and FI can be found in the Supplementary Materials.

**Joint Representations (CODIRI and CODIFI).** While isovist representations can encode significantly more information than image representations, they are limited by visibility, meaning that image representations can provide key information unseen by isovists. The proposed joint image-isovist representations $\mathbf{E}' \in \mathbb{R}^{61 \times n' \times n'}$ concatenate image and isovist representations along the channel dimension, taking advantage of their respective strengths. In particular, we favor CODI as the image part of the joint representation since it encodes more information than BODI, and we use either RI or FI for the isovist part since they are complementary to each other with respect to the geometry that they encode well. This results in the joint representations CODIRI and CODIFI, which concatenate CODI with RI and FI respectively (Figure 1).

## 5 Modeling

We first investigate how several existing convolutional encoder-decoder (CED) networks perform LTCFP when adapted to the proposed super-resolution framework: SegNet [Badrinarayanan *et al.*, 2017], U-Net [Ronneberger *et al.*, 2015], Attention U-Net [Oktay *et al.*, 2018]. Since the these models were designed for image segmentation, we first replace the final softmax layer of each model with a $1 \times 1$ convolution layer followed by a sigmoid layer for regression. Then, in order for these models to predict a $m \times m$ output ($m = 256$) from a $n' \times n'$ input ($n' = 64$), we add two up-convolution blocks to the decoder prior to the final layer. Each up-convolution block consists of a transposed convolution layer followed by 2 sequences of convolutional, batch normalization, and ReLU activation layers. The transposed convolution layers in these blocks use $4 \times 4$ kernels with stride $= 2$ and padding $= 1$, and all other convolution layers utilize $3 \times 3$ kernels with stride $= 1$ and padding $= 1$.

### 5.1 GeoInteractNet (GINet)

Although using additional up-convolution blocks will inherently produce higher-resolution predictions, the feature maps that the existing CED networks normally use to feed high-resolution spatial information to their decoders are missing for the additional blocks, hindering the learning of fine-grained details in super-resolution. Therefore, we propose GeoInteractNet (GINet), a model coupling a Multi-SCale Attention Network (M-SCAN) and a CED prediction network in parallel (Figure 2). M-SCAN first feeds $\mathbf{X}$ into a Geodesic Interaction Module to generate a super-resolution geodesic interaction map $\mathbf{I} \in \mathbb{R}^{1 \times m \times m}$, which approximates human interaction along the shortest paths to the goal (i.e., geodesic paths). For each individual in the crowd $\mathbf{A}$, the Geodesic Interaction Module computes the geodesic path along $\mathbf{G}$ and encodes it as a $m \times m$ binary path image, in which pixels along the geodesic path are 1. $\mathbf{I}$ is then computed as the normalized sum of all path images, where values close to 1 indicate the potential for interactions between individuals.

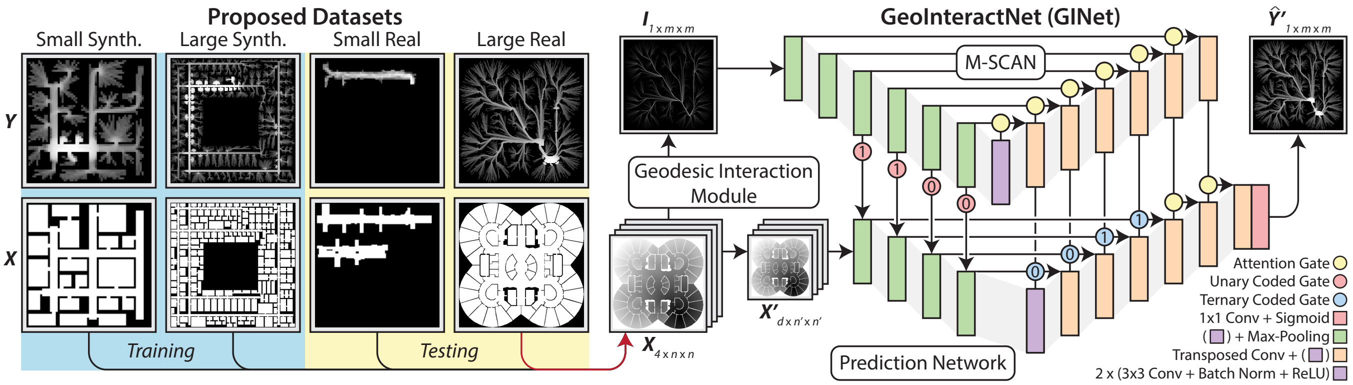

Figure 2: **GINet.** Our proposed model GINet uses a Geodesic Interaction Module and Multi-SCale Attention Network (M-SCAN) to produce super-resolution priors Prediction Network's decoder blocks, allowing GINet to predict fine-grained details better than existing models. We have generated two synthetic datasets to train our models and two real environment and crowd datasets exclusively for testing.

M-SCAN then inputs $\mathbf{I}$ into the Attention U-Net architecture [Oktay *et al.*, 2018]. The architecture of the prediction network parallels that of M-SCAN, but the first two encoder blocks are removed in order to accept $\mathbf{X}'$ as input (Figure 2). In order to pass information from M-SCAN to the prediction network, we use a mix of attention gates [Oktay *et al.*, 2018], unary coded gates, and ternary coded gates. The first 4 encoder blocks in the prediction network are connected to matching encoder blocks in M-SCAN through unary coded gates $\upsilon$ (Figure 2).

$$\upsilon(\mathbf{M}_e, c) = c\mathbf{W}\mathbf{M}_e,$$

where $\mathbf{M}_e$ is the $b$-channel output from M-SCAN's encoder block, $\mathbf{W} \in \mathbb{R}^{b \times 1}$ is a $1 \times 1$ convolution, and $c \in \{0, 1\}$ is a hyperparameter that determines whether the gate is enabled. The output of $\upsilon$ is concatenated to the input of the prediction network's encoder block. The first 4 decoder blocks in the prediction network upscale their outputs $\mathbf{P}_d$ to match the sizes of corresponding encoder blocks $\mathbf{P}_e$ in the prediction network and decoder blocks $\mathbf{M}_d$ in M-SCAN. All three blocks' outputs have $b$ channels and are passed to a ternary coded gate $\tau$ (Figure 2).

$$\tau(\mathbf{M}_d, \mathbf{P}_e, \mathbf{P}_d, c) = \sigma_2\Big(\mathbf{W}_4\sigma_1\big(\mathbf{W}_1\mathbf{P}_d + c\mathbf{W}_2\mathbf{M}_d + (1-c)\mathbf{W}_3\mathbf{P}_e\big)\Big),$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{b \times b/2}$ and $\mathbf{W}_4 \in \mathbb{R}^{b/2 \times 1}$ are a $1 \times 1$ convolutions, $\sigma_1$ is a ReLU activation, $\sigma_2$ is a sigmoid activation, and $c \in \{0, 1\}$ is a hyperparameter that determines whether the gate uses feature maps from the prediction network's encoder or M-SCAN's decoder. The 1-channel result of $\tau$ is duplicated to $b$ channels, element-wise multiplied with $\mathbf{P}_e$, and then concatenated to $\mathbf{P}_e$ as input to the next decoder block in the prediction network. The final 2 decoder blocks in the prediction network use Attention U-Net's attention gates. However, since there are not matching encoder blocks in the prediction network, the attention gates use matching decoder blocks from M-SCAN (Figure 2).

In sequential order, the unary and ternary coded gates can be hyperparameterized by a byte $\mathbf{c} \in \{0, 1\}^8$, where the $i$th coded gate is parameterized by the bit $\mathbf{c}_i$. We investigate 10 variants of GINet, e.g., GINet$_{[00001111]}$, which disables all unary coded gates and enables all ternary coded gates.

## 6 Experimental Preliminaries

LTCFP is most suited for quickly assessing the risk of overcrowding in situations with large crowds and environments, meaning that training LTCFP models inherently requires ground truth data with overcrowding. There are obvious logistical and ethical concerns that prohibit the acquisition of such data with real humans at a large enough scale to use for training. Therefore, we rely on the same solution as practitioners in application domains [Ma and Yarlagadda, 2015], which is to simulate crowds using the Social Force model [Helbing and Molnar, 1995]. The same model was used by the only existing LTCFP dataset [Sohn *et al.*, 2020], but our datasets include more varied synthetic environments, larger simulated crowds, and large real environments.

**Small and Large Synthetic Environment Datasets.** The 2 synthetic datasets consists of 8,000 total training and 2,400 total testing crowd scenarios with thousands of unique synthetic environments. In order to represent the variety of real environments, we have extended the procedural method from [Sohn *et al.*, 2020] to generate 48 unique types of environments, featuring different combinations of 2 sizes ($64 \times 64$ and $224 \times 224$ m$^2$), 3 types of exterior shapes, 4 types of corridor arrangements, and either curved or axis-aligned geometry. The size determines whether the environment is in the small or large synthetic dataset. Each generated environment was used to simulate two crowd scenarios with a low (5%) and high (20%) chance of initializing an agent for each navigable cell in the environment, which affects the amount of overcrowding. Both scenario types tasked their crowds with navigating to a single random goal location, and after every agent reached the goal, the ground truth crowd flow map was computed.

**Small and Large Real Environment Datasets.** The large real dataset contains 12 floorplans of environments that vary in terms of function (e.g., railway stations, hospitals, and shopping centers) and locality (e.g., Austria, China, and Germany). Each environment was scaled to $224 \times 224$ m$^2$ and used to simulate 100 crowd scenarios at 5% occupancy, yielding a total of 1,200 crowd scenarios. The small real dataset consists of 1,500 crowd scenarios produced using real human trajectories from two regions in a $64 \times 64$ m$^2$ train terminal [Alahi *et al.*, 2014], where each scenario involves at least 30 people simultaneously moving toward the same location. We *do not consider this as an LTCFP dataset* because its environment and crowds are too small to require LTCFP in practice. The small real dataset's crowds have tens of real humans while the other datasets have hundreds or thousands of agents. However, this dataset strikes a balance between the needs of HTP models and LTCFP models, making it useful for evaluating the whether HTP models can be used for LTCFP. Given the small number of real environments, we reserve both real datasets for testing only.

**Training Protocol.** Adam optimization [Kingma and Ba, 2014] was used for training U-Net, Attention U-Net, and GINet, while stochastic gradient descent was used for CAGE and SegNet (with momentum $= 0.9$). Prior to training, the data was shuffled, and the batch size was set to 4. All models were then trained exclusively on both synthetic datasets for 100 epochs with a learning rate of $0.01$ from $\{0.1, 0.01, 0.001\}$, which performed best across models. The loss function was set to Mean Absolute Error (MAE), which performed better than Mean Squared Error and Binary Cross-Entropy Loss. After training, these models were tested on the synthetic datasets and both real datasets, which were never seen during training. A machine with an Intel Core i9-9960X 3.10 GHz, 64GB RAM, and an NVIDIA GeForce RTX 2080 Ti 11GB was used for all training and testing.

**Evaluation Protocol.** We report the MAE between the ground truth $\mathbf{Y}$ and predicted crowd flow $\widehat{\mathbf{Y}}$. Crowd flow maps are visualized as heatmaps and predictions are accompanied by a difference image $\mathbf{D} = \widehat{\mathbf{Y}} - \mathbf{Y}$, where 0 values are white, positive values (overestimation) become increasingly red and negative values (underestimation) become increasingly blue.

## 7 Experiments

In this section, we report the results of three experiments to determine (1) whether a SOTA HTP model can perform LTCFP, (2) how well the proposed representations, the super-resolution framework, and GINet perform against the SOTA CAGE framework, and (3) which variant of the proposed GINet model is best.

### 7.1 LTCFP using Human Trajectory Prediction (HTP)

Recently, an evaluation of SOTA HTP models showed that the predicted trajectories of pedestrians often collide with each other and move through obstacles [Sohn *et al.*, 2021]. HTP models are often able to ignore these issues, because they are evaluated on the basis of the minimum error across $k$ short-term samples. However, these issues become exacerbated when HTP models are adapted to LTCFP, which forces them to commit to one sample and use it as input for the subsequent prediction.

We have evaluated Trajectron++ (T++) [Salzmann *et al.*, 2020], a SOTA HTP model, according to this method (with $k = 1$) on the Small Real dataset. After training on 1,500 crowd scenarios, Trajectron++ was primed with real input trajectories and made to predict $\sim$38 seconds into the future, i.e., the average duration of a trajectory in the dataset. The resulting trajectories for all agents were then converted into a crowd flow map and compared with predictions made by the CAGE framework. The qualitative results in the Supplementary Materials show that trajectories almost completely ignore the obstacles in the environment despite making its predictions conditioned on the environment. We have also masked the crowd flow map with the environment map to boost its performance, but the quantitative results in Table 1 indicate that even with masking, Trajectron++ performs significantly worse than both CAGE (the baseline) and all other LTCFP models on a dataset of real crowds in a small real environment (which is normally suitable for HTP). Although Attention U-Net outperforms GINet, the Small Real dataset is not a proper LTCFP dataset, meaning that it should be ignored when evaluating GINet with other LTCFP models.

| **Model** | T++ | T++<br>(Masked) | CAGE | SegNet | U-Net | Attention<br>U-Net | GINet<br>[11110000] |
|---|---|---|---|---|---|---|---|
| **Input** | N/A | N/A | BODI | CODIFI | CODIFI | CODIRI | CODI |
| **MAE** | 0.09868 | 0.02668 | 0.01607 | 0.01296 | 0.01005 | **0.00999** | 0.01024 |

Table 1: This table compares Trajectron++ (T++), a SOTA HTP model, adapted to LTCFP with LTCFP models using MAE.

| **Model** | | **CAGE<br>(SOTA)** | **SegNet** | **U-Net** | **Attention<br>U-Net** | **GINet**<br>[11110000] |
|---|---|---|---|---|---|---|
| **Small Synth.** | Best Input | CODI | RI | BODI | CODI | CODIFI |
| | Best MAE | 0.03955 | 0.02942 | 0.01476 | 0.01533 | **0.01399** |
| | BODI MAE | 0.04182 | 0.02971 | 0.01476 | 0.01627 | 0.01403 |
| | $\Delta\%$ MAE | 5.4280 | 0.9761 | 0.0000 | 5.7775 | 0.2851 |
| **Large Synth.** | Best Input | CODIFI | CODIRI | CODIRI | CODIFI | CODIFI |
| | Best MAE | 0.08026 | 0.04777 | 0.03065 | 0.03158 | **0.0212** |
| | BODI MAE | 0.08694 | 0.06028 | 0.03166 | 0.03293 | 0.02184 |
| | $\Delta\%$ MAE | 7.6835 | 20.7532 | 3.1901 | 4.0996 | 2.9304 |
| **Large Real** | Best Input | CODIFI | FI | CODIRI | CODIRI | CODI |
| | Best MAE | 0.06495 | 0.06109 | 0.04098 | 0.04194 | **0.01983** |
| | BODI MAE | 0.07023 | 0.06751 | 0.04986 | 0.05014 | 0.02046 |
| | $\Delta\%$ MAE | 7.5182 | 9.5097 | 17.8099 | 16.3542 | 3.0792 |

Table 2: The above table compares CAGE (the SOTA) with image segmentation networks adapted to LTCFP and our proposed model GINet across columns. Within each column, BODI (the SOTA representation) is compared with proposed representations using the percent difference ($\Delta\%$) in MAE.

### 7.2 Model and Representation Comparison

Table 2 compares the CAGE framework, CED networks adapted to the proposed super-resolution framework, and GINet after training on all input representations, where BODI is the SOTA baseline. For each model, the table reports the best input representation, the MAE for that input, the MAE when using BODI, and the percent improvement in MAE from BODI to the best input representation. Across all models from CAGE to GINet, we find that BODI is rarely the best representation and that joint representations (CODIRI and CODIFI) often lead to better MAE. The average percent improvement that the best input has over BODI across

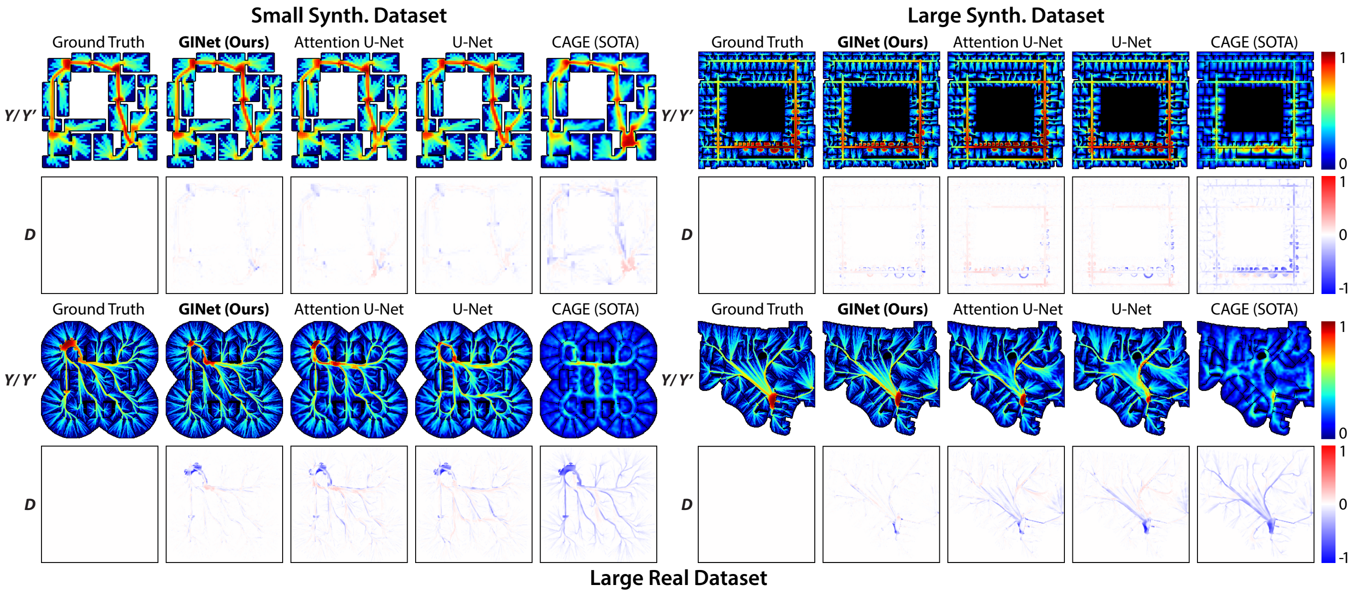

Figure 3: This figure compares the ground truth with prediction images from CAGE, U-Net, SegNet, and GINet using the difference image $\mathbf{D}$. We highly recommend zooming in to see the nuanced, but important differences between GINet and other models.

all models and datasets is ∼14.1%. Between different models using their best representations, it appears that the CAGE framework performs worse on every dataset than all proposed models by a substantial margin. On the other hand, the GINet$_{[11110000]}$ variant shows the best MAE across all LTCFP datasets (which excludes the Small Real dataset). Its MAE on the Large Real dataset never seen during testing is *∼51.6% lower* than the next best model, and its MAE on the Large Synthetic dataset is *∼30.8% lower* than the next best model, which clearly indicates that GINet scales well to large and complex environments. Figure 3 visualizes predictions made for each dataset. The full evaluation of each input representation and qualitative results have been provided in the Supplementary Materials.

### 7.3 GINet Variant Comparison

GINet$_{[11110000]}$ was chosen for comparison with other models after an ablation study was conducted over GINet variants using `CODIFI`. Table 3 shows the MAE for each variant. In general, it appears that having all coded gates enabled is worse than having them all disabled. This does not include the final two attention gates, which are enough to greatly improve performance on large datasets compared to adapted CED networks (see GINet$_{[00000000]}$). It also appears that the geodesic interaction map serves better as attention than solely as input with $\mathbf{X}'$, because GINet$_{[11110000]}$ performs better than GINet$_{[11000000]}$. Interestingly, GINet sees the most benefit from *repeatedly* receiving multi-scale geodesic interaction maps as input to its encoder blocks.

| **Model** | **Datasets** | | |
|---|---|---|---|
| | Small Synth. | Large Synth. | Large Real |
| GINet$_{[00000000]}$ | 0.01557 | 0.02376 | 0.02211 |
| GINet$_{[11111111]}$ | 0.02171 | 0.02701 | 0.02966 |
| GINet$_{[11000000]}$ | 0.01583 | 0.02244 | 0.02113 |
| GINet$_{[00110000]}$ | 0.01558 | 0.02365 | 0.02223 |
| GINet$_{[00001100]}$ | 0.01552 | 0.02242 | 0.02108 |
| GINet$_{[00000011]}$ | 0.01950 | 0.02591 | 0.02208 |
| GINet$_{[11110000]}$ | **0.01399** | **0.02120** | **0.02018** |
| GINet$_{[00111100]}$ | 0.01559 | 0.02219 | 0.0209 |
| GINet$_{[00001111]}$ | 0.02203 | 0.02675 | 0.02454 |
| GINet$_{[11000011]}$ | 0.01780 | 0.02591 | 0.02342 |

Table 3: This table reports an ablation study of GINet variants trained using `CODIFI`.

## 8 Conclusion

In this work, we have proposed (1) novel input representations using egocentric visibility information, which improve both the SOTA and several more performant models, (2) a super-resolution framework, which shows that CED networks for image classification can be repurposed for LTCFP and perform much better than the SOTA, (3) GINet, which shows that geodesic interaction is the key to improving LTCFP, and (4) a dataset of real and synthetic environments and crowds that is far more suitable for evaluating the scalability of models to large and complex environment than the only existing LTCFP dataset.

## Acknowledgements

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 22STESE00001 01 01.

## References

[Alahi *et al.*, 2014] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014.

[Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[Benedikt, 1979] Michael L Benedikt. To take hold of space: isovists and isovist fields. *Environment and Planning B: Planning and design*, 6(1):47–65, 1979.

[Bohannon, 2005] John Bohannon. Directing the herd: Crowds and the science of evacuation. *Science*, 310(5746):219–221, 2005.

[DESA, 2018] UN DESA. 68% of the world population projected to live in urban areas by 2050, says un. *United Nafions Department of Economic and Social Affairs*, 2018.

[Gupta *et al.*, 2018] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[Helbing and Molnar, 1995] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kuhl and Giardina, 1982] Frank P Kuhl and Charles R Giardina. Elliptic fourier features of a closed contour. *Computer graphics and image processing*, 18(3):236–258, 1982.

[Ma and Yarlagadda, 2015] Wenbo Ma and Prasad KDV Yarlagadda. Pedestrian dynamics in real and simulated world. *Journal of Urban Planning and Development*, 141(3):04014030, 2015.

[Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Salzmann *et al.*, 2020] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. *arXiv preprint arXiv:2001.03093*, 2020.

[Sewall *et al.*, 2010] Jason Sewall, David Wilkie, Paul Merrell, and Ming C Lin. Continuum traffic simulation. In *Computer Graphics Forum*, volume 29, pages 439–448. Wiley Online Library, 2010.

[Sohn *et al.*, 2020] Samuel S Sohn, Honglu Zhou, Seonghyeon Moon, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. Laying the foundations of deep long-term crowd flow prediction. In *European Conference on Computer Vision*, pages 711–728. Springer, 2020.

[Sohn *et al.*, 2021] Samuel S. Sohn, Mihee Lee, Seonghyeon Moon, Gang Qiao, Muhammad Usman, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. A2x: An agent and environment interaction benchmark for multimodal human trajectory prediction. *Motion, Interaction and Games*, 2021.

[Turner and Penn, 2002] Alasdair Turner and Alan Penn. Encoding natural movement as an agent-based system: an investigation into human pedestrian behaviour in the built environment. *Environment and planning B: Planning and Design*, 29(4):473–490, 2002.

[Van Den Berg *et al.*, 2011] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-Body Collision Avoidance. In *Robotics Research*, pages 3–19. Springer, 2011.

[Wiener and Franz, 2004] Jan M Wiener and Gerald Franz. Isovists as a means to predict spatial experience and behavior. In *International Conference on Spatial Cognition*, pages 42–57. Springer, 2004.

[Wiener *et al.*, 2007] Jan M Wiener, Gerald Franz, Nicole Rossmanith, Andreas Reichelt, Hanspeter A Mallot, and Heinrich H Bülthoff. Isovist analysis captures properties of space relevant for locomotion and experience. *Perception*, 36(7):1066–1083, 2007.

[Zhang *et al.*, 2002] Dengsheng Zhang, Guojun Lu, et al. A comparative study of fourier descriptors for shape representation and retrieval. In *Proc. 5th Asian Conference on Computer Vision*, page 35. Citeseer, 2002.

[Zhen *et al.*, 2008] Wang Zhen, Liu Mao, and Zhao Yuan. Analysis of trample disaster and a case study–mihong bridge fatality in china in 2004. *Safety Science*, 46(8):1255–1270, 2008.