remote sensing

MDPI

Article

LSN-GTDA: Learning Symmetrical Network via Global Thermal Diffusion Analysis for Pedestrian Trajectory Prediction in Unmanned Aerial Vehicle Scenarios

Ling Mei [1,2], Mingyu Fu [1], Bingjie Wang [2], Lvxiang Jia [2], Mingyu Yu [1], Yu Zhang [1] and Lijun Zhang [3,*]

[1] School of Electronic Information, Wuhan University of Science and Technology, Wuhan 430081, China; meil3@wust.edu.cn (L.M.); fumy@wust.edu.cn (M.F.); ymy@wust.edu.cn (M.Y.); zhangy@wust.edu.cn (Y.Z.)
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; bjwang@hust.edu.cn (B.W.); lvxiangjia@hust.edu.cn (L.J.)
[3] Department of Electrical and Electronic Engineering, Department of Mechanical and Mechatronics Engineering, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
* Correspondence: lzhang@sun.ac.za

Abstract: The integration of pedestrian movement analysis with Unmanned Aerial Vehicle (UAV)-based remote sensing enables comprehensive monitoring and a deeper understanding of human dynamics within urban environments, thereby facilitating the optimization of urban planning and public safety strategies. However, human behavior inherently involves uncertainty, particularly in the prediction of pedestrian trajectories. A major challenge lies in modeling the multimodal nature of these trajectories, including varying paths and targets. Current methods often lack a theoretical framework capable of fully addressing the multimodal uncertainty inherent in trajectory predictions. To tackle this, we propose a novel approach that models uncertainty from two distinct perspectives: (1) the behavioral factor, which reflects historical motion patterns of pedestrians, and (2) the stochastic factor, which accounts for the inherent randomness in future trajectories. To this end, we introduce a global framework named LSN-GTDA, which consists of a pair of symmetrical U-Net networks. This framework symmetrically distributes the semantic segmentation and trajectory prediction modules, enhancing the overall functionality of the network. Additionally, we propose a novel thermal diffusion process, based on signal and system theory, which manages uncertainty by utilizing the full response and providing interpretability to the network. Experimental results demonstrate that the LSN-GTDA method outperforms state-of-the-art approaches on benchmark datasets such as SDD and ETH-UCY, validating its effectiveness in addressing the multimodal uncertainty of pedestrian trajectory prediction.

Keywords: UAV; trajectory prediction; global thermal diffusion; complete response analysis; multi-modality

Academic Editor: Massimiliano Pepe

Received: 16 November 2024
Revised: 24 December 2024
Accepted: 27 December 2024
Published: 4 January 2025

Citation: Mei, L.; Fu, M.; Wang, B.; Jia, L.; Yu, M.; Zhang, Y.; Zhang, L. LSN-GTDA: Learning Symmetrical Network via Global Thermal Diffusion Analysis for Pedestrian Trajectory Prediction in Unmanned Aerial Vehicle Scenarios. *Remote Sens.* **2025**, *17*, 154. https://doi.org/10.3390/rs17010154

1. Introduction

Remote sensing research plays a crucial role in various human analysis fields, such as detection [1], tracking [2], path planning [3,4], and group re-identification [5] using UAV aerial imagery data. The highD dataset demonstrates that UAVs are essential for capturing the trajectories of traffic participants, including humans [6]. UAVs offer the advantage of recording naturalistic behavior in large-scale environments through camera-equipped drones. Pedestrian movement analysis and UAV-based remote sensing research

are closely intertwined, as both fields leverage spatiotemporal data to enhance the understanding of human movement patterns and environmental interactions from an elevated perspective. The former enriches the latter by providing ground-level insights that improve the calibration and interpretation of aerial data, ultimately enhancing the precision of spatial analyses [7–9]. The data collected from UAVs are critical for predicting human motion, which is of great importance for self-driving vehicles in addressing intelligent obstacle avoidance challenges posed by surrounding moving objects, such as autonomous robots [10,11], pedestrians [12–14], and vehicles [15,16].

Human movement is inherently stochastic, influenced by a variety of factors, including individual behaviors and environmental conditions, e.g., a latent decision for a long-term goal or a random decision against environmental changes. Thus, the future trajectory of pedestrians could be influenced by behavioral and stochastic factors; the former includes social interactions (e.g., avoiding collisions, following others, or walking together), and the latter includes scene semantics (e.g., walls, barriers, or traffic), dynamic obstacles (e.g., moving vehicles, cyclists, or other pedestrians), etc. This complex phenomenon introduces a multimodal challenge, marked by significant uncertainty. As a result, pedestrians can exhibit a range of plausible future trajectories with multimodal predictions.

However, humans are target-oriented agents who actively express their intentions through actions to achieve desired targets. Therefore, analyzing this uncertainty requires a comprehensive understanding of pedestrian movement patterns in the context of environmental factors. The multimodal uncertainty may bring about the influence of cognitive consciousness on pedestrian motion, avoidance of interactions with others, or sharp turns by other individuals that lead to changes in the path. However, existing methods fail to adequately address and model this multimodal uncertainty within a solid theoretical framework.

Motivated by the long-term target-oriented principles underlying human motion, we integrate historical trajectories and scene semantics into our approach. The inherent uncertainty in trajectory prediction is addressed through a binary classification framework, which combines potential long-term goals inferred from the historical behaviors of pedestrians, as well as environmental factors and path selection variables influenced by their habits. Specifically, this framework considers habitual factors derived from pedestrian historical trajectories and stochastic decision factors arising from environmental conditions and individual preferences. Therefore, the multimodal uncertainty analysis of trajectory targets extends the predictive horizon for pedestrian trajectories.

On the other hand, since the pedestrian trajectory prediction is time-dependent, we use the energy diffusion principle to explain its uncertainty. The future distribution of particles is considered in a thermodynamic framework. Thus, under high-uncertainty conditions, particles (positions) are randomly distributed across all walkable regions. While under low-uncertainty conditions, particles aggregate and deform into a clear trajectory.

Due to the temporal correlation of the pedestrian trajectory prediction, we propose employing the energy diffusion principle to elucidate its inherent uncertainty. In our conceptual framework, we analogize future pedestrian positions to particles within the domain of thermodynamics. Therefore, our approach aims to learn from this diffusion process by gradually decomposing the uncertainty of trajectory prediction into interpretable factors and converting the ambiguous prediction regions into diverse deterministic trajectories of multi-modality. Then, we design a global and comprehensible deep learning network to fuse the information from humans and their surrounding environment and factorize the multimodal problem into some explicit factors to analyze. For the framework of deep learning networks, we use the U-Net network [17] as the backbone sub-network, as it is conducive to semantic segmentation. Due to the symmetrical structure of the U-Net, our

designed overall network also exhibits symmetry, which facilitates the global representation and interpretability of the trajectory encoding and decoding sub-networks.

In this paper, we propose a symmetrical U-Net architecture to learn both the stochastic and behavioral factors in pedestrian trajectory prediction through global thermal diffusion analysis. First, we perform semantic segmentation on the scene map to identify walkable areas, which will serve as the stochastic factor in our prediction framework, representing the environmental input within our prediction framework. Next, one U-Net branch is employed to learn semantic information, identifying plausible path nodes and target points. We argue that, for a given scene map, the potential movement targets and path nodes are static and unaffected by the agent's historical movements. Therefore, for the path node and target encoder, we use only the processed scene graph as input, which significantly improves training efficiency (as demonstrated in Section 4.6). In parallel, an additional U-Net branch is used to decompose the behavioral factor by learning from past trajectories. Specifically, the trajectory branch integrates both historical trajectory data and scene information. This is achieved by combining the processed scene map with the trajectory heatmap, which is then fed into the trajectory encoder. Furthermore, the output from the node and target heatmaps helps the trajectory decoder make a global estimation of the trajectory distribution for a future time.

Therefore, we design a novel neural network architecture named LSN-GTDA by symmetrical U-Net [17] sub-networks, which not only reduces the estimating error significantly but also markedly reduces the training time needed to achieve a convergent model. There are two input ports in the LSN-GTDA framework corresponding to the scene heatmap encoder and the trajectory encoder, respectively.

In addition, to factorize the inherent uncertainty of human trajectory into some explicit factors, we apply a novel signal and system-based thermal diffusion process using a complete response mechanism in the proposed symmetrical network. Specifically, we refer to the energy diffusion of the complete response mechanism, which can analyze the historical part and the stochastic part in the future in terms of human movements. Hence, if we take the start of the predicting time horizon as the zero-moment, the diverse motion patterns arising from the two uncertainty factors correlate with the corresponding parts of the complete response mechanism in the field of signal and system, then they can be analyzed by the energy diffusion principle based on the complete response. Therefore, we can use the input of the semantic information and historical trajectory from the symmetrical network to conduct the diffusion process. Through the above analysis, the operation of the symmetrical network is explicable.

In total, the contributions in this paper are threefold and are listed as follows:

- A novel symmetrical U-Net-based framework is proposed, which integrates pedestrian historical trajectories with scene semantic segmentation information. Through a bottom–up global analysis of trajectory nodes, paths, and targets, the framework enables multimodal long-term predictions that leverage historical patterns and account for future stochastic uncertainties.
- A global thermal diffusion mechanism is introduced for the symmetrical network, utilizing the global concept of a complete response mechanism. This approach enhances the interpretability of the model, addressing uncertainty within the network and providing a faster inference efficiency and deeper understanding of the underlying principles of human trajectory prediction.
- Extensive experiments are conducted on human trajectory prediction in diverse UAV scenarios, demonstrating competitive accuracy and efficiency. The results show that our approach outperforms state-of-the-art baselines.

2. Related Works

2.1. Research on Human Trajectory in UAV Scenarios

In previous work, to find the motion state of humans in UAV scenarios, researchers have undertaken significant efforts to create effective goal-driven models to anticipate trajectory distribution such as Y-Net [18], SGNet [19], and PECNet [20]. These methods handle the uncertainty of prediction tasks as a multi-modality model. To capture the implicit clues in the multi-modality problem, some related works use Social LSTM [21] and Social GAN [22] to abstract pedestrians as points. On the other hand, to utilize more motion clues, Liang et al. [23] designed a multi-task learning network to predict human motion by using the surrounding interaction information of human behavior. Karttikeya et al. [18] proposed the Y-Net method based on an analysis of multiple modalities to extend the time horizon from five seconds to one minute.

Researchers have often implemented predictions of pedestrian trajectories utilizing inference. These inferences are usually based on the historical trajectories of pedestrians [24] and a given RGB static scene map [22,25–27]. Previous works have used the premise that pedestrian movements are goal-oriented [18,28]. The current academic view is that the trajectory of pedestrian movement is a mixture of certainty and uncertainty. Certainty emerges from the constraints imposed by the specific scenario in which it resides, whereas uncertainty can be influenced by subjective pedestrian factors or external factors such as unconscious randomness [29]. Since the pedestrian's motion is flexible, they still have multiple feasible future paths to walk. To obtain multiple predictions, some studies in the literature [25,30] applied Generative Adversarial Networks (GANs) to generate diverse trajectories. Some methods [31,32] used Conditional Variational Autoencoder (CVAE) and KL divergence to train the networks, and Maeda et al. [33] used flow-based generative methods to obtain multimodal results. In general, it appears that multi-modality pedestrian trajectory prediction is a problem that researchers are currently enthusiastic to solve. However, despite the relentless pursuit of prediction accuracy and time duration through intricate networks, existing works overlook the exploration of a theory that could uncover the inherent mechanism underlying the issue of human trajectory prediction.

2.2. Pedestrian Trajectory Model Simulation

UAVs are often used to record traffic scenarios since they are usually equipped with high-definition cameras and can capture a broad view from the air [6]; how to predict the position and direction of movement of the target on the road has become an important topic. However, pedestrian trajectory prediction under the UAV's view is a complex task that involves predicting the future movement of pedestrians in a crowd [34]. To achieve this, scientists have developed various methods for learning pedestrian behavior and parsing the motion patterns. For example, Yue et al. [35] proposed a trajectory prediction model to explain pedestrian behavior by surrounding social correlation. Wang et al. [19] proposed a recurrent network to explore pedestrian behavior for trajectory estimation. The Fourier Transform, accompanied by spectral analysis, is employed to capture pedestrian behavior [36]. To promote coherent motion patterns within crowds, diffusion-based methodologies are utilized to strengthen the correlations between adjacent particles [37–40]. In addition, the social interaction around pedestrians is also used to design a unified trajectory prediction model based on the attention mechanism [41]. Wong et al. [42] proposed an angle-based trainable social interaction representation named SocialCircle to reflect the correlation with surrounding agents along with trajectories. Kim et al. [43] proposed a relational reasoning graph-based method to predict the human's trajectory by using high-order social interactions. These methods simulate each waypoint in a trajectory as an

individual crowd member with coherent motion patterns according to the analysis of their behavioral trajectory.

Recently, to simulate the trajectory model further, Davis et al. [44] proposed a guided diffusion model to constrain trajectories by learning the surrounding environment context. Mao et al. [45] used a trainable leapfrog strategy to design a diffusion-based model named LED to skip some denoising steps in the generation and accelerate inference speed during trajectory prediction.

However, the diffusion process is applied to specific locations and executed individually for each trajectory, which does not effectively reduce the overall number of reverse diffusion steps. Additionally, these diffusion models require annotating a large amount of ground truth data for future trajectories based on past observations. Since the behavioral trajectory only reflects the pedestrian's past state, it motivates us to predict future trajectories by integrating aleatoric uncertainty with behavioral factors through global thermal diffusion analysis. Moreover, the significant computational cost associated with diffusion models hinders their use for real-time prediction. Therefore, we propose an alternative approach that replaces multiple denoising steps with a more efficient response mechanism to achieve faster inference speeds.

3. The Proposed Method

Generally, as shown in Figure 1, the main uncertainty affecting the human trajectory prediction can be factorized into the past behavioral influence and the future stochastic factors, which is a multi-modality problem. Therefore, we design the human trajectory predicting framework as follows.

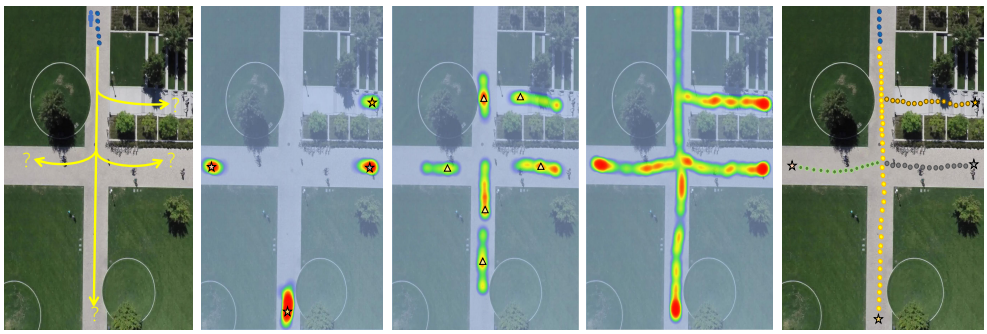


Figure 1. Illustration of the research process for pedestrian trajectory prediction from a UAV perspective. (**a**) Multi-modality of the trajectory prediction; (**b**) the behavioral factor in the prediction over the target nodes of the zero-input response; (**c**) the stochastic factor over the path nodes of the zero-state response; (**d**) the thermal distribution in the prediction; (**e**) each color indicates predicted trajectories for different target modality. The pentacle and triangle symbols mean the targets and nodes in a trajectory, respectively.

Let $\mathbf{b}_k = (x_k, y_k) \in \mathbb{R}^2$ denote the 2D coordinates of a pedestrian in the image scene, and the sequence of the observed past positions is denoted by $\{\mathbf{b}_k\}_{k=1}^{k_p}$ during the past period of $T_p = k_p / T_{fps}$ seconds, where T_{fps} defines the frame rate. LSN-GTDA is designed to forecast positions of the human in the subsequent temporal period, e.g., the position $\{\mathbf{b}_k\}_{k=k_p+1}^{k_p+k_f}$ in the next T_f seconds, where $T_f = k_f / T_{fps}$.

The future predictions of human motion are uncertain. Pedestrian trajectory prediction involves multiple possibilities; there are multiple patterns for both the destination and the optional path, which makes the prediction a multi-modality problem. Hence, we divide the overall uncertainty into two factors: the behavioral factor and the stochastic factor. We define the number of multiple predictions of the final trajectory target as M_b, which

depends on the behavioral uncertainty. In addition, the number of multiple predictions of the path access to a target is defined as M_s, which reflects the stochastic uncertainty. In the short-term prediction case, M_s is usually small, and M_b is more dominant since the path diversity is limited by the short time horizon; e.g., the total predicted trajectories are the same as M_b if $M_s = 1$. In the long-term prediction case, both of them dominate since there is more uncertainty in the long-term horizon, and moderate diversity of the predicted path ($M_s > 1$) can reduce the uncertainty, which will be evaluated in Section 4.5.1.

3.1. Thermal Diffusion Process

To analyze the proposed framework in Figure 2, we divide this network into two parts: the processing of past trajectories of pedestrians and scene maps. The former can be regarded as the system's input due to its strong variability; i.e., the historical trajectory of each pedestrian can vary greatly. The latter, on the other hand, can be considered the inherent state of the system due to its spatial invariance, where the path points and target points generated in the same scene remain unchanged, or it remains invariant due to the change of the input. We consider the current prediction moment to be the zero-moment, and the system's output is our prediction for the pedestrian trajectory.

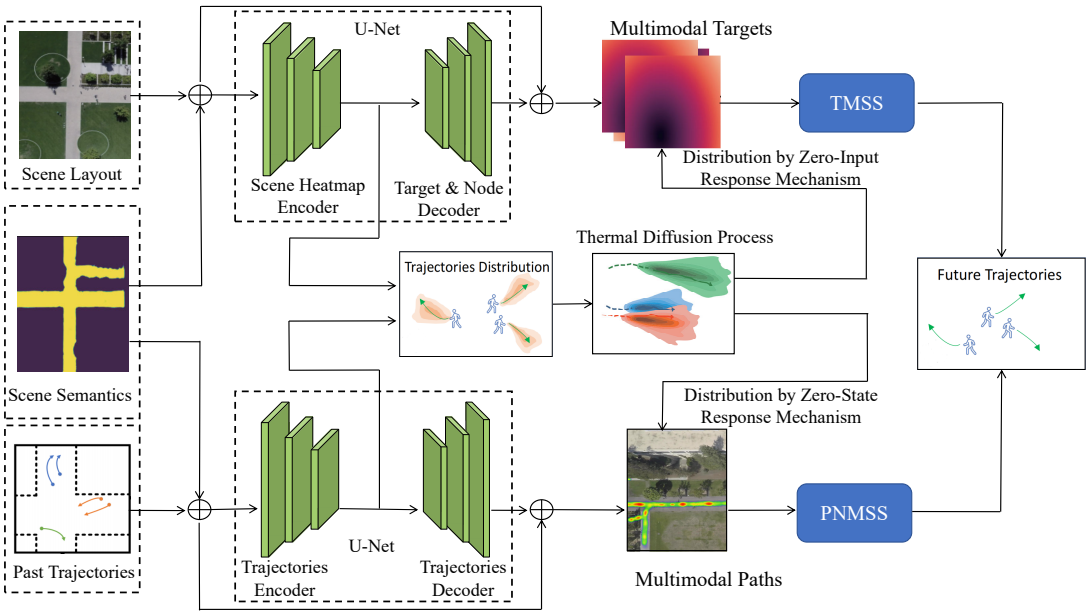


Figure 2. The framework of the proposed LSN-GTDA pedestrian trajectory prediction method. LSN-GTDA comprises a scene segmentation module and a trajectory heatmap module, which constitutes symmetrical U-Net architectures including both target and trajectory branches. The decoding output uses the global thermal diffusion process including zero-input and zero-state response to predict the future trajectory, and TMSS and PNMSS are used to handle the target and path diversity of multimodality, respectively.

Human trajectories exhibit collective movement patterns by a series of waypoints, which reflect the behavior of human gaits. To find precise and coherent motion patterns within the trajectory, the optical flow field [46] can be used to construct the global motion correlation of the waypoints in the trajectory. Specifically, each waypoint is regarded as a heat source node and is connected with neighboring nodes through energy diffusion in the optical flow field [47], which models the inherent relationship between these waypoints of a trajectory.

3.1.1. Behavior Consistency in Neighboring Nodes of the Trajectory

At first, according to the neighboring similarity in [34], the similarity between a pair of neighboring path nodes i and j can be defined as

$$S_t(i,j) = \max(R_t(i,j), 0) \tag{1}$$

where $R_t(i,j)$ measures the correlation of velocity between pair-wise nodes at time t. $R_t(i,j) \in (0,1)$ reflects the consistent behavior of an individual among its neighbors. To further describe the correlation for the waypoints that are not adjacent, we propose a graph structure by the connectivity of its associating nodes.

Let $\mathbf{M}$ be the weighted adjacency matrix to describe the correlation in a graph, which is used to associate with the node set Γ of the trajectory. The similarity $R_t(i,j)$ is an edge between the pair-wise nodes in Equation (1). Let $\eta_h = \{w_0 \to w_1 \to \cdots \to w_h\}$ denote a trajectory with length h by nodes $w_0, w_1, \cdots, w_h$ on $\mathbf{M}$ between individual nodes 0 and h. Then, the probability of the real trajectory for path η_h can be defined as

$$g_{\eta_h} = \prod_{b=0}^{h} R_t(w_b, w_{b+1}) \tag{2}$$

Since there might be more than one path between the node i and j, let the set ψ_h contain all paths of length h between them; then, the h-path similarity is represented as

$$g_h(i,j) = \sum_{\eta_h \in \psi_h} g_{\eta_h}(i,j) \tag{3}$$

According to Theorem 1 in [34], $g_h(i,j)$ can be efficiently computed by matrix $\mathbf{M}$ since it is the (i,j) entry of $\mathbf{M}^h$.

3.1.2. The Thermal Diffusion Process by an Analysis of the Complete Response Mechanism

Based on the above discussions, a further description of the thermal diffusion process among the nodes of the path will be described in this section. Since the h-path similarity in Equation (3) measures the behavioral consistency of pair-wise nodes, each node in the path can be treated as a "heat source" that diffuses energy to other nodes; thus, we can refer to a classical idea from [48] that models the thermal diffusion as follows:

$$\frac{\partial \mathbf{E}_{\mathbf{P}_i,t}}{\partial t} = m_p^2 \left(\frac{\partial^2 \mathbf{E}_{\mathbf{P}_i,t}}{\partial x^2} + \frac{\partial^2 \mathbf{E}_{\mathbf{P}_i,t}}{\partial y^2}\right) + \mathbf{F}_{\mathbf{P}_i} \tag{4}$$

where $\mathbf{E}_{\mathbf{P}_i,t} = [E^x_{\mathbf{P}_i,t}, E^y_{\mathbf{P}_i,t}]$ is the thermal energy for the particle at node $\mathbf{P}_i = (p_i^x, p_i^y)$ after performing the propagation of thermal energy for t seconds, and m_p is the propagation coefficient. $\mathbf{F}_{\mathbf{P}_i} = [f^x_{\mathbf{P}_i}, f^y_{\mathbf{P}_i}]$ is the input motion vector for node $\mathbf{P}_i$. Without $\mathbf{P}_i$, Equation (4) can be solved by [37] as follows:

$$\mathbf{E}_{\mathbf{P}_i,t} = \frac{1}{HW} \sum_{\mathbf{P}_j \in \xi, \mathbf{P}_j \neq \mathbf{P}_i} e^{\alpha}_{\mathbf{P}_i,t}(\mathbf{P}_j) \tag{5}$$

where $\mathbf{E}_{\mathbf{P}_i,t}$ is the final diffused thermal energy for the waypoint after t seconds, ξ is the set of all waypoints in the trajectory, and H and W are the height and width of the input image. Motivated by the complete response theory from the discipline of Signals and Systems [49], the first term in Equation (4) models the diffusion of thermal energies over free space so that the spatial correlation among path nodes can be properly enhanced during energy propagation, which can be viewed as the zero-state response if the initial

state of the thermal diffusion process is supposed to be none at the zero-moment. As [37] mentions, after T seconds, the spreading energy from $\mathbf{P}_i$ to $\mathbf{P}_j$ is

$$e^{\alpha}_{\mathbf{P}_i,T}(\mathbf{P}_j) = u^{\alpha}_{\mathbf{P}_j} \times e^{-m_p \left\|\mathbf{P}_i - \mathbf{P}_j\right\|^2} \times e^{m_f \left\|\mathbf{F}_{\mathbf{P}_j} \cdot (\mathbf{P}_i - \mathbf{P}_j)\right\|^2} \tag{6}$$

where m_p and m_f are the propagation coefficient and the force propagation factor, respectively. $\alpha \in (x, y)$ denotes the axis. $\mathbf{F}_{\mathbf{P}_j}$ is the optical flow that describes the motion tendency at position $\mathbf{P}_j$, $\mathbf{U}_{\mathbf{P}_j} = (u^x_{\mathbf{P}_j}, u^y_{\mathbf{P}_j})$ is the current motion pattern of $\mathbf{P}_j$, which is initialized by $\mathbf{U}_{\mathbf{P}_j} = \mathbf{U}_{\mathbf{P}_i}$ and its thermal energy in location j is $\left|\mathbf{U}_{\mathbf{P}_j}\right|$ [37]. However, the energy diffusion strategy of [37] ignores the initial motion state $\mathbf{F}_{\mathbf{P}_i}$; due to the lack of consideration of the initial state before the pair-wise particles' energy diffusion, i.e., the zero input state, its mechanism is too limited to handle the global prediction issue. Therefore, we introduce a global formulation that adds the zero-input response to the thermal diffusion process. Let $E_{\mathbf{P}_i}$ be the total motion energy at location $\mathbf{P}_i$; it consists of two parts. One is the zero-state response $E^{zs}_{\mathbf{P}_i} = \sqrt{(e^x_{\mathbf{P}_i})^2 + (e^y_{\mathbf{P}_i})^2}$, and the other is the zero-input response $E^{zi}_{\mathbf{P}_i}$. We consider the pedestrian's current position as the zero-moment in the global system analysis to predict the subsequent waypoints they may pass; $E^{zs}_{\mathbf{P}_i}$ represents the uncertainty of the stochastic factor in the future, whereas $E^{zi}_{\mathbf{P}_i}$ considers the epistemic factor before the zero-moment. Mathematically,

$$E^{zs}_{\mathbf{P}_i} = \left|\mathbf{U}_{\mathbf{P}_i}\right| \times e^{-m_p} \times e^{m'_f \left|\mathbf{F}_{\mathbf{P}_i}\right|} \tag{7}$$

$$E^{zi}_{\mathbf{P}_i} = \left|\mathbf{U}_{\mathbf{P}_i}\right| \times e^{m_i \left|\mathbf{F}_{\mathbf{P}_i}\right|} \times \cos(\mathbf{F}_{\mathbf{P}_i}, \mathbf{F}_{\mathbf{P}_j}) \tag{8}$$

Equation (7) is obtained by simplifying Equation (6). Since i and j are adjacent nodes, we assume the length of $\left\|\mathbf{P}_i - \mathbf{P}_j\right\| = 1$ in Equation (6), m'_f and m_i are two coefficients. To consider the real road environment, the background scenario limits some roads, which can be passed, while some parts cannot priorly, which is the zero-input response that indicates the historical information such as previous trajectories. On the other hand, the zero-state response focuses on the stochastic factor including the random motion choice of the pedestrian by the background map after the zero-moment. According to Equations (2) and (3), the energy associated with each node signifies the likelihood of pedestrian passage, thereby influencing the heatmap distribution of the subsequent adjacent node through thermal diffusion. Greater energy levels indicate a higher likelihood of passage. The optimal path can be inferred through the integration of the global thermal diffusion process outlined above.

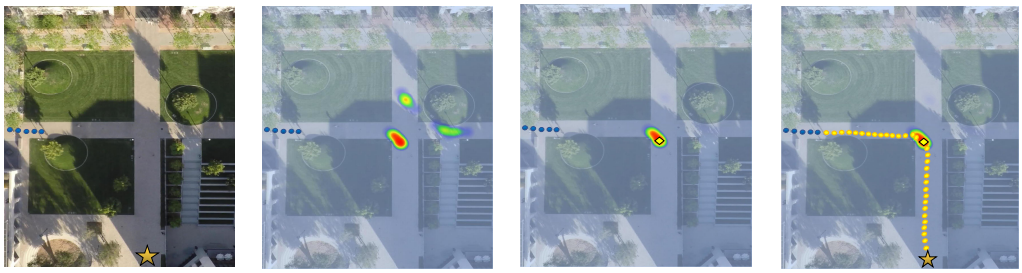
We factorize the whole uncertainty into two explicit parts, which can be analyzed by the two components of the complete response mechanism correspondingly. In this practical approach, categorizing the uncertainties within the model helps identify which uncertainties can potentially be reduced, thereby enhancing the interpretability of the prediction study. Hence, to validate the interpretability of the proposed method, we analyze the effectiveness of the two proposed uncertainty factors: the stochastic factor and the behavioral factor. We propose the "Target Modality Sampling Strategy" (TMSS) to give a specific number of target samples needed for evaluation. TMSS investigates the impact of varying modality sampling numbers on the outcomes within a specified target quantity framework, which reflects the behavioral factor with the zero-input response in the proposed thermal diffusion process. TMSS uses the K-means method to directly control the sampled modalities and the number of cluster centers. Therefore, we use TMSS, which is cognizant of the number of the target samples, M_b, needed in the evaluation procedure.

In addition, the target and path are dependent on each other; we use another trick named "Path and Node Modality Sampling Strategy" (PNMSS) to analyze the function of

the stochastic factor by different path modalities for a given target such as Equation (3), which reflects the impact of zero-state response in the proposed thermal diffusion process.

Therefore, it is useful to build the complete response to describe the thermal diffusion model by zero-input and zero-state response, which provides theoretic interpretability for the design of the following proposed deep learning-based network.

Figure 3 shows visual representations of TMSS and PNMSS; the target and the path node are dependent on each other, where the road forks into three different paths. If the sampled target is located at the bottom, the pedestrian tends to move downwards rather than upwards in the map, which results in not passing the upper waypoint in Figure 3b.



(a) (b) (c) (d)

Figure 3. Visualization of the proposed LSN-GTDA pedestrian trajectory prediction method on SDD. (**a**) Historical path nodes and the motion target marked as a yellow star; (**b**) diverse waypoint distribution; (**c**) resulting waypoint distribution; (**d**) the predicted trajectory result to the goal.

According to the above analysis, we propose a hierarchical prior information to model the process. After the targets are sampled, we fix the target to constrain the possible positions of the next path node. We assume the node lies on a straight line segment between the sampled target and the past trajectory. Then, we use a multivariate Gaussian prior with means at the assumed position to relax the assumption, which is multiplied pixel-wise to the distribution of the predicted path node. The thermal diffusion is represented by the distribution. Fusing the prior and the predicted distribution can allow for the obtention of scenario-compliant path nodes in a feasible direction. Finally, we use softargmax operation [50] for the first node and sample the remaining nodes randomly to obtain 2D points from the distribution. We repeat the above process for the next node to realize the thermal diffusion in the path generation.

In summary, as shown in Figure 4, TMSS provides possible motion targets by using the existing historical motion state, which represents the zero-input response in the diffusion process. PNMSS provides feasible paths for a specific target by using diverse input factors regardless of the initial state after the zero-moment, which represents the zero-state response in the diffusion process. The proposed thermal process combines them for the purposes of global analysis to consider the diversity of the target and the path.

3.2. Symmetrical LSN-GTDA Human Trajectory Forecasting Structure

As shown in Figure 2, LSN-GTDA is a symmetrical network that comprises two U-Net-based branches. The above branch analyzes the behavioral uncertainty by learning the semantic information of input scenes to estimate the optional target distribution, and feeds together with the scene heatmap encoder to the nether branch to enhance the decoding of the trajectory. The nether branch analyzes the stochastic uncertainty by learning from the past trajectory and walkable area for prediction. We further use the zero-input response and the zero-state response to estimate the heatmap distribution for the two branches, respectively. Finally, we predict the global future trajectory from the heatmap.

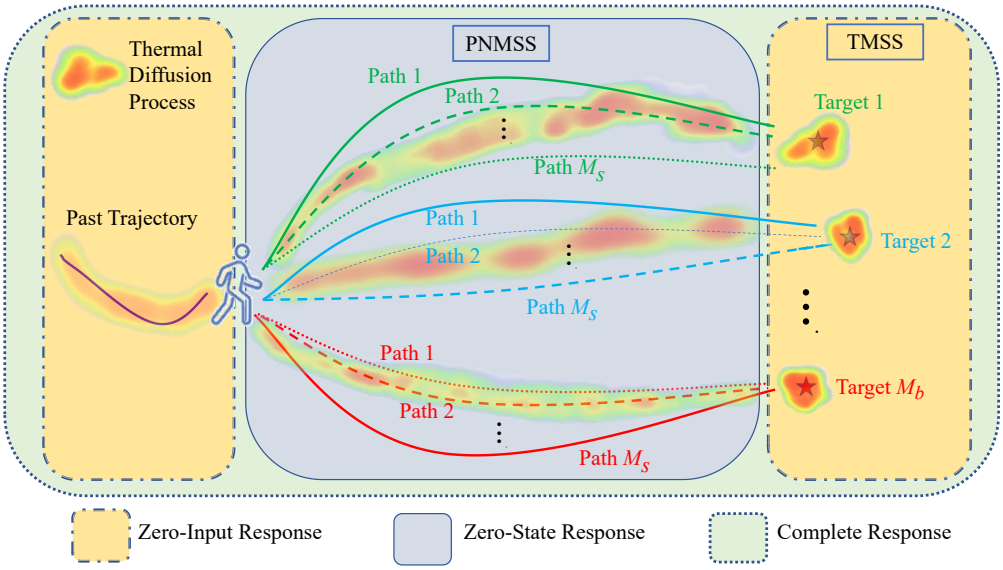


Figure 4. A structural diagram of the TMSS and PNMSS strategy in the proposed global thermal diffusion process. Different colors and lines denote diverse prediction modalities.

Generally, trajectory points may be randomly distributed across the entire space under high uncertainty. By using the proposed global thermal diffusion process, the diffused global energy can make the predicted trajectory more focused through its heatmap and thus reduce uncertainty.

To understand the scene of the pedestrian effectively, semantic information is exploited as an aided and useful clue together with the trajectory prediction module. Therefore, we formulate the deep learning network of human trajectory prediction as two modules: One is the semantic segmentation module, and the other is the trajectory forecasting module.

For the semantic segmentation module, we use a segmentation network by U-Net to obtain a segmentation map $\mathbf{G}_I$ of scene image I, which includes N_c classes of behaviors such as standing, walking, running, etc. On the other hand, the previous action history $\{\mathbf{b}_k\}_{k=1}^{k_p}$ of the agent h is transformed to a trajectory heatmap $\mathbf{Q}$ of k_p channels that comprise spatial sizes I; they correspond to the past T_p seconds sampled at the frame rate. The trajectory heatmap is defined as

$$\mathbf{Q}(k,i,j) = \frac{\|(i,j) - \mathbf{b}_k\|}{\max\|(x,y) - \mathbf{b}_k\|}, (x,y) \in I \tag{9}$$

To add interactions and constraints such as avoiding obstacles for the agents with the given scene, we pretrain the semantic segmentation model to use the sparse scene image data efficiently. The U-Net model [17] and ResNet101 [51] backbone are used in our implementation. The ResNet101 encoder's weights are pretrained on ImageNet, while the weights for the segmentation head and U-Net decoder are randomly initialized. The images are downsampled four times and padded to be divisible by 32 as required for U-Net; then, the size is cropped to 256 × 256 pixels. The SDD segmentation model is trained using the ADAM optimizer.

The semantic segmentation module is crucial for understanding the environment in which pedestrians navigate. By segmenting the scene into different regions based on semantic categories (e.g., sidewalks, roads, crosswalks, buildings), this segmentation helps the model distinguish between walkable and non-walkable areas, enabling more realistic predictions of pedestrian movement.

Then, we concatenate the trajectory heatmap with the semantic map, which represents the probability distribution of where a pedestrian is likely to move in the future. Different

from [18], the proposed trajectory feature and semantic feature are encoded independently. Therefore, the channel dimensions of the semantic encoder and the trajectory encoder are $H \times W$ and $k_p + N_c$, respectively.

The semantic segmentation module informs the trajectory heatmaps by providing a contextual understanding of the environment. These two components create a dynamic and context-aware framework for predicting pedestrian movement. While the semantic segmentation module ensures that the prediction is grounded in the real-world environment, the trajectory heatmaps enable the model to predict not just the location but the future path distribution, capturing the inherent uncertainty in human motion.

In this way, the semantic information can be introduced along with the channel dimension to produce semantic heatmap representation, which is passed to the encoder as its first input channel. The overall network is symmetrical in terms of the structure of the encoder and decoder parts.

3.3. Encoder of the Trajectory Module

In terms of the trajectory module, we consider two aspects of its function. First of all, the observed scene heatmap tensor $\mathbf{Q_G}$ is processed with an additional U-Net encoder $\mathbf{E}$ rather than a common U-Net encoder that is shared by the semantic module and the trajectory module. In this way, the U-Net encoder [17] consists of $N_\mathbf{E}$ blocks with max pooling and increased channel depth correspondingly. In this way, the final deep representation $\mathbf{Q_E}$ and $N_\mathbf{E} - 1$ intermediate feature tensors are spatially compact, and various spatial resolutions are provided to the subsequent goal decoder.

Moreover, as shown in Figure 2, the proposed trajectory module is connected with the output of the semantic segmentation decoder $\mathbf{D}$, which propagates useful semantic information to the proposed trajectory module and then assists the module in paying more attention to the region that contains goals that are more possible to achieve, while ignoring other unimportant regions. Pedestrians are always drawn to destinations; therefore, we use goal attraction from additional semantic information in order to guide the trajectory to the goal.

3.4. Decoder

3.4.1. Target and Node Heatmap Decoder

The main role of this decoder is to generate useful clues about the target and path node for human motion. As we have raised the size of the feature map through the $N_\mathbf{E}$ blocks in the encoder, many details are unavoidably lost during the convolution and pooling operations of the encoder. Thus, we need to merge the symmetry features from the LSN-GTDA encoder, which can bring apparent recovery in the resolution. Hence, they are combined as an input terminal to the trajectory decoder, and we use the expansion arm of the scene heatmap encoder as a reference for the decoder. Then, the heatmap provides a visual way to represent possible targets by mapping predicted probabilities or densities for the trajectory decoder, which handles the uncertainty in multi-modality predictions. In total, the heatmap illustrates uncertainty and variability, while target nodes provide concrete predictions.

3.4.2. Trajectory Decoder

The trajectory decoder is similar to the target and node heatmap decoder. The difference is that we also need to gather the information on node and target points with the branch of the segmentation map and past trajectory heatmaps. Thus, an extra input of the U-Net's expansion arm [17] is added, which is the output of the Node and Target Point Decoder. To fit the diverse resolutions, we down-sample the node heatmaps for different decoder layer requirements to match the size of the input information. After the final

per-pixel sigmoid processing, we obtain a probability estimation of human trajectory in the next T_f seconds.

3.5. Loss Function

Due to the trajectory being used in scene representation, we design losses to reflect the discrepancy between the estimated probability distribution and the ground truth (GT). Since human motion is inherently goal-directed, it exerts its will by actions to realize a desired effect [28]. Therefore, a Gaussian heatmap can be predetermined based on previous probabilistic models such as prior variance, which can reduce the unnecessary interference caused by the uncertainty of the goal through the inherent goal-driven property of humans. GT is modeled as a Gaussian heatmap $\mathbb{Q}$, which is centered at the observed locations with a predefined variance. To avoid the misguidance caused by an inappropriate variance, it is chosen adaptively by the human's last observed position and the sampled multimodal target; the ground truth $\mathbb{Q}$ also represents the energy distribution according to the thermal diffusion process in Equation (4). The energy increases with probability and decreases with increasing uncertainty. We use the ground truth to train the Node and Target Decoder by designing the corresponding loss function.

Since KL divergence is capable of measuring the differences between various probability distributions and real distribution, and given that the pedestrian trajectory shows continuity in the 2D scene, it is not appropriate to use the cross-entropy loss function that is suitable for discrete classification. Instead, we utilize KL divergence to construct the loss function, which consists of the losses between the real distributions $\mathbb{Q}(\cdot)$ and the estimated distributions $\widetilde{\mathbb{Q}}(\cdot)$ of the target, node, path distributions, and diffusion process, as follows:

$$L_{\text{target}} = KL(\mathbb{Q}(\mathbf{b}_{k_p+k_f}), \widetilde{\mathbb{Q}}(\mathbf{b}_{k_p+k_f})) \tag{10}$$

$$L_{\text{node}} = \sum_{b=1}^{h} KL(\mathbb{Q}(\omega_b), \widetilde{\mathbb{Q}}(\omega_b)) \tag{11}$$

$$L_{\text{path}} = \sum_{\eta_h \in \psi_h} KL(\mathbb{Q}(\eta_h), \widetilde{\mathbb{Q}}(\eta_h)) \tag{12}$$

$$L_{\text{diffusion}} = \sum_{\mathbf{P}_i \in \xi} KL(\mathbb{Q}(\mathbf{E}_{\mathbf{P}_i,t}), \widetilde{\mathbb{Q}}(\mathbf{E}_{\mathbf{P}_i,t})) \tag{13}$$

Then, we obtain the overall loss function according to the above equations:

$$L = \alpha_1 L_{\text{target}} + \alpha_2 L_{\text{node}} + \alpha_3 L_{\text{path}} + \alpha_4 L_{\text{diffusion}} \tag{14}$$

where α_1, α_2, α_3, and α_4 control the trade-off between different loss terms.

4. Experiments

To validate the performance of the proposed method, we use two benchmarks to compare LSN-GTDA with other state-of-the-art methods, e.g., the Standford Drone Dataset (SDD) and the ETH-UCY Dataset. The former is used to evaluate both short-term and long-term predictions in UAV scenarios, while the latter is mainly used to evaluate short-term predictions.

4.1. Datasets

Standford Drone Dataset (SDD): The dataset includes more than 11,000 independent pedestrians across 20 top–down scenes captured on the Stanford University campus in bird's eye view using a flying UAV. There are over 40,000 agent–scene interactions in the dataset, and it has been widely used in the trajectory prediction literature in short temporal horizon settings. For short-term prediction, to show consistency with the other baselines, we follow the same setting with the Y-Net [18] baseline to make comparisons, with samples at FPS = 2.5 to obtain an input sequence of length $k_p = 8$ (3.2 s) and output of length

$k_f = 12$ (4.8 s). Moreover, for long-term settings, the raw dataset is split in the same fashion as proposed in the TrajNet benchmark [52], evaluating the same scenes, all of which are not seen during the training stage. The long-term setting downsamples the data to FPS = 1 and obtains an input sequence of $t_p = 5$ s and output of length $t_f = 30$ s.

ETH-UCY Dataset: Many previous methods such as [21,22] use several sub-datasets from ETH [53] and UCY [54] to train and evaluate their models with the "leave-one-out" strategy, which is called the ETH-UCY dataset. It contains 1536 pedestrians with five different scenes, including thousands of non-linear trajectories and annotations of the pedestrians' positions in meters. We follow the parameter setting that is the same as that of SDD in the experiment of the ETH-UCY dataset. The ETH-UCY dataset is a widely used benchmark for pedestrian trajectory prediction, consisting of several sequences collected in urban environments. Here is a brief introduction to each of the sequences: (1) ZARA1 is collected in a busy street in Zurich, which captures pedestrians in a densely populated area, showcasing a variety of walking patterns and interactions. The camera is stationary. (2) ZARA2 describes another area in Zurich, similar to ZARA1 but with different environmental settings; it adds more variety in terms of the scene and pedestrian density. (3) UNIV records human movement on the university campus in Cyprus, including students and staff moving in and out of university buildings, which exhibits a mix of pedestrian interactions. (4) HOTEL records pedestrians moving towards and away from a hotel in Zurich, highlighting behaviors in a more controlled environment with varying pedestrian densities. It includes various interactions, such as people waiting and entering the hotel. (5) ETH captures pedestrian traffic around the university's main entrance of the ETH Zurich campus. It features various paths, including crossing streets and moving towards entry points to show the interactions and movements of pedestrians.

4.2. Evaluation Metrics

In the experiment, both the established Average Displacement Error (ADE) and Final Displacement Error (FDE) are used as the metrics to evaluate the performance of future forecasting. The former reflects the average ℓ_2 error between the forecasted future and the ground truth over the entire trajectory, while the latter calculates the ℓ_2 error between the forecasted future and the ground truth for the final predicted point [55]. In terms of multiple future predictions, we follow the prior works [18,22] to report the final error as the minimum error over all predicted future scenarios. Specifically, ADE and FDE are defined as follows:

$$\mathrm{ADE}(\mathbf{b}^h_{-k_p:k_f}) = \frac{\sum_{h=1}^{N}\sum_{k=1}^{k_p} ||\mathbf{b}^h_k - \hat{\mathbf{b}}^h_k||^2}{N \times k_p} \tag{15}$$

$$\mathrm{FDE}(\mathbf{b}^h_{-k_p:k_f}) = \frac{\sum_{h=1}^{N} ||\mathbf{b}^h_{k_p} - \hat{\mathbf{b}}^h_{k_p}||^2}{N} \tag{16}$$

where $\mathbf{b}^h_{-k_p:k_f}$ defines the previous positions during past k_p frames and the predicted positions during subsequent k_f frames of the $h-$th pedestrian, which constitutes the predicted trajectory. N is the number of predicted persons, and $\mathbf{b}^h_k$ and $\hat{\mathbf{b}}^h_k$ are the ground truth and predicted position, respectively.

4.3. Short-Term Predicting Results

4.3.1. Results of Stanford Drone Dataset

As shown in Table 1, we follow the setting of [18] to make fair comparisons, and the results of short-term predicting are presented with the setting of $t_p = 3.2$ s and $t_f = 4.8$ s. Since there are limited random probabilities in the short-term predicting case, we set

$M_s = 1, M_b = 5$ and compare the proposed LSN-GTDA method with other state-of-the-art baselines. Table 1 shows our method for realizing an ADE of 5.98 and an FDE of 6.80, which improves the previous strong baseline such as the NSP-SFM [35] method by 8.3% on ADE and 35.9% on FDE. It also outperforms some latest state-of-the-art baselines, e.g., E-V^2-Net-SC [42] by 8.6% on ADE and 34.4% on FDE. Previous works [35,42] are usually limited by context awareness and the absence of interaction modeling, while the proposed thermal diffusion-based method captures multimodal uncertainty by not only considering the trajectories of neighboring nodes but also incorporating environmental factors.

Table 1. Short temporal horizon predicting results on SDD. The label "*" of HighGraph* denotes the baseline of "EigenTrajectory [56] + HighGraph" shown in [43]. "W" means the conference's workshop. All the bold font method denotes the best results in this paper.

Method	Venue/Year	ADE	FDE
S-GAN [22]	CVPR 2018	27.23	41.44
CF-VAE [57]	NeurIPS(W) 2019	12.60	22.30
P2TIRL [58]	Arxiv 2020	12.58	22.07
SimAug [59]	ECCV 2020	10.27	19.71
PECNet [20]	ECCV 2020	9.96	15.88
Y-Net [18]	ICCV 2021	7.85	11.85
NSP-SFM [35]	ECCV 2022	6.52	10.61
MID [60]	CVPR 2022	7.61	14.30
LED [45]	CVPR 2023	8.48	11.66
TUTR [41]	ICCV 2023	7.76	12.69
HighGraph* [43]	CVPR 2024	7.81	11.09
MlgtNet [61]	TITS 2024	6.91	11.04
E-V^2-Net-SC [42]	CVPR 2024	6.54	10.36
LSN-GTDA(Ours)	-	**5.98**	**6.80**

For the sake of comparison with published studies about the trajectory prediction on the diffusion process, e.g., MID [60] and LED [45], the results of Table 1 have validated that the proposed LSN-GTDA outperforms them with a large margin, which validates the superiority of the proposed global thermal diffusion analysis. Moreover, since LSN-GTDA uses the uncertainty factorization and complete response mechanism rather than a denoise-based diffusion process, it overcomes the limitation of expensive time consumption compared with MID. The inference time analysis is shown in Section 4.6.

4.3.2. Results of the ETH-UCY Dataset

In addition, to validate the generalization of the proposed method, we also report the visualization results of ETH-UCY in Figure 5, the predictions in different scenarios of ETH-UCY are quite consistent with the ground truth, which is much better than the state-of-the-art baseline, i.e., E-V^2-Net-SC [42]. Although the scenarios in the ETH-UCY dataset are complex, the proposed LSN-GTDA estimates the motion tendency under diverse environments.

4.4. Long-Term Predicting Results

To evaluate the effect of *behavioral* and *stochastic* uncertainty, we propose a long-term trajectory forecasting setting with a stable prediction duration of 60 s in Figure 6, which is much longer than other prior baselines. Table 2 reports our results on the Standford Drone dataset (SDD) for a time horizon of $t_f = 30$ s of prediction in the future with $t_p = 5$ s input. $t_f = 30$ s means the temporal midway between the observed inputs and the longest estimated goal in Figure 6. The long-term results of LSN-GTDA are at the $M_s = 5, M_b = 50$ setting for a fair comparison with other methods in Table 2 and Figure 6. R-PECNet is trained by a recurrent short-term PECNet model [20]. On the Stanford Drone Dataset (SDD), the proposed LSN-GTDA method outperforms the state-of-the-art method on the long horizon setting as well; e.g., it improves Y-Net in terms of ADE/FDE performance from

47.94/66.71 to 31.91/41.45 with a large margin (33.4%/37.9%). Furthermore, the proposed LSN-GTDA method promotes the performance of PECNet by decreasing ADE/FDE by 55.8%/64.9%.

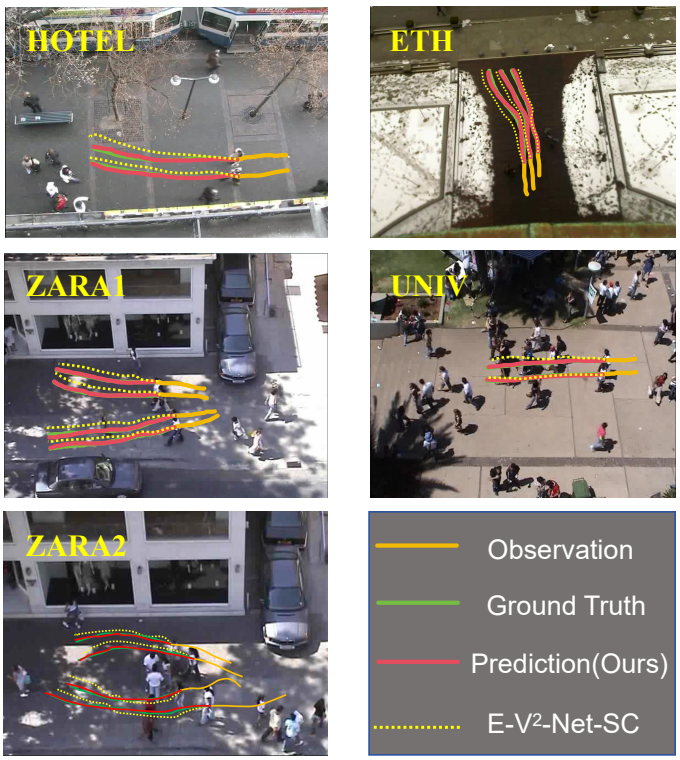


Figure 5. Visualization of predicted trajectories compared with the state-of-the-art on the ETH-UCY dataset.

Moreover, we promote the prediction horizons to one minute and observe the performance results in Figure 6. The ADE errors of all methods increase with the prolongation of prediction time, and the performance of the proposed LSN-GTDA method and the Y-Net method are significantly better than PECNet in terms of long-term prediction, which indicates the importance of building target and node models in long-term trajectory predictions. In addition, Figure 6 also shows that the proposed LSN-GTDA method outperforms the Y-Net method.

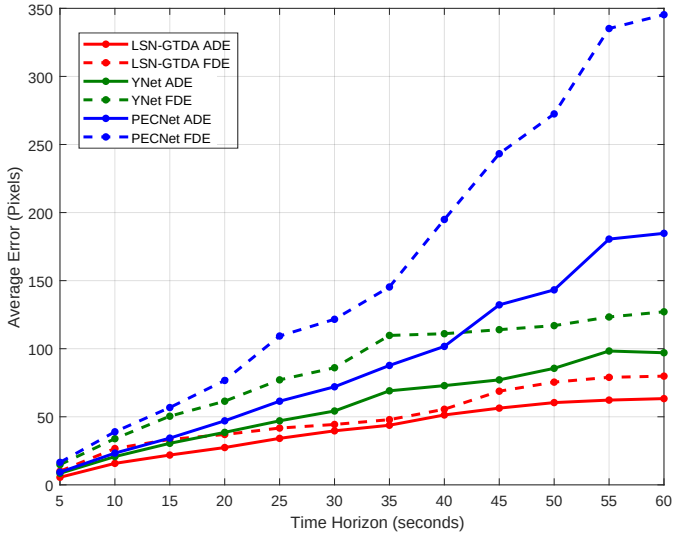


Figure 6. Benchmarking performance against time horizons.

Table 2. Long-term trajectory predicting results on SDD benchmark. We compare LSN-GTDA with 4 long-term baselines: S-GAN (B1), PECNet (B2), R-PECNet (B3), and Y-Net (B4). The bold font signifies the best results.

Method	B1 [22]	B2 [20]	B3 [20]	B4 [18]	LSN-GTDA (Ours)		
$\mathbf{M_s}$	-	-	-	-	1	2	5
ADE	155.32	72.22	261.27	47.94	39.68	37.26	**31.91**
FDE	307.88	118.13	750.42	66.71	44.36	41.79	**41.45**

4.5. Ablation Study of the Proposed Thermal Diffusion Process

To conduct the ablation study for the proposed thermal diffusion process, we use the ADE and FDE metrics to carry out the evaluation—firstly, in situations where several samples are required, such as during the evaluating procedure, sampling from the estimated distribution $\widetilde{\mathbb{Q}}(\cdot)$. However, this approach may ignore other sub-optimal ones such as samples from some redundant adjacent regions or low-probability regions, which leads to multimodal deficiency. Hence, the proposed TMSS is used to evaluate the influence of the number of target samples, which reflects the role of the zero-input response in the diffusion process. In addition, PNMSS is used to evaluate the influence of path multimodality on the already sampled targets, which reflects the role of the zero-state response in the diffusion process.

4.5.1. Analysis of TMSS for Zero-Input Response

To evaluate the performance of TMSS in the proposed thermal diffusion process, we fix M_s to observe the evolution of M_b in Figure 7. For a given M_s, ADE tends to decrease with the increase in M_b, indicating that M_b can help M_s improve the performance, which validates the necessity of behavioral uncertainty M_b. For the target multi-modality parameter M_b given by past historical trajectory states, the diversification of path multi-modality parameter M_s depending on future random factors can reduce ADE errors to some extent. This phenomenon indicates that the zero-input response sets feasible goals through TMSS according to the tendency of the past trajectory; thus, the above results of variable M_b validates the finding that TMSS reasonably provides diverse targets for subsequent multimodal path selection in the thermal diffusion process.

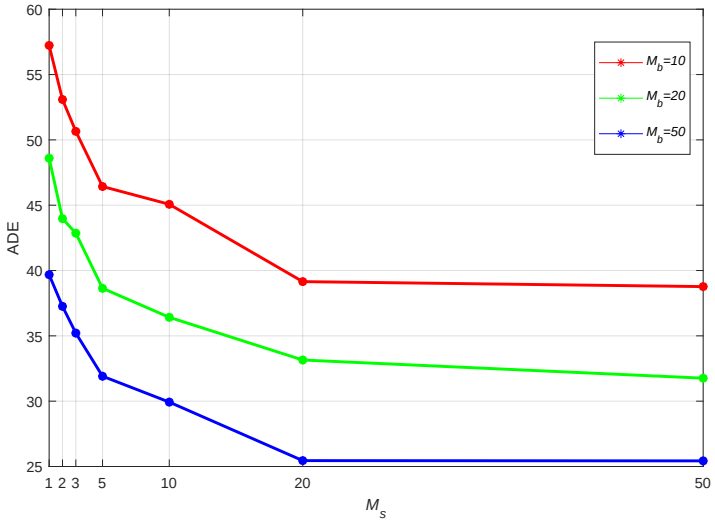


Figure 7. Evolution performance of two multimodal uncertainty parameters for the proposed LSN-GTDA on the SDD long-term benchmark. We fix the amount of the target modality (M_b) to observe the effect of the multi-modality path.

Moreover, a fast K-means clustering approach is used to set the number of clusters to $M_b - 1$; all cluster centers are generated by the K-means approach, along with the softargmax sampled point. In this way, the TMSS strategy controls the sampled points directly, which does not need the "truncation trick" proposed in PECNet [20]. Figure 3 shows the visualization result of the proposed method; both tabular and visual results validate the effectiveness of the proposed complete response-based thermal diffusion mechanism in our method.

4.5.2. Analysis of PNMSS for Zero-State Response

In addition to the analysis of M_b, to further explore the role of zero-state response in the proposed thermal diffusion process, we also report the results of different M_s on the SDD dataset to observe its evolution performance in Figure 7, which shows that ADE decreases as M_s increases for a given M_b, and indicates the effective use of multi-modality stochastic parameters.

Specifically, for each curve with a fixed M_b in Figure 7, consistent improvement is observed in ADE with increasing M_s, indicating the effectiveness of the diversity of PNMSS, which depends on stochastic factors from new inputs after the zero-moment (prediction start), which corresponds to the zero-state response generated by the random state input after the zero moment in the proposed thermal diffusion process. Moreover, Table 2 reports the LSN-GTDA's performance compared with different baselines by using different M_s; we observe that our proposed method has improved significantly in ADE on the SDD dataset, which indicates that the multi-modality of paths can reduce the error of estimating the same final target. Hence, the results in Table 2 also validate the finding that a moderate gain of path modality (M_s) can reduce prediction errors in long-term cases.

To further evaluate the components of the thermal diffusion process in our method, we conduct an additional ablation study on the SDD benchmark. The results shown in Table 3 validate the effectiveness of long-term prediction with both TMSS and PNMSS strategies; "None (LSN-GTDA)" indicates that the proposed LSN-GTDA does not use the thermal diffusion process, and "TMSS (LSN-GTDA)", "PNMSS (LSN-GTDA)", and "TMSS+PNMSS (LSN-GTDA)" represent the use of the respective module and whole parts of the thermal diffusion process, respectively. The results in Table 3 show that the whole parts perform much better than others in long-term cases, which demonstrates the effectiveness of the proposed global thermal diffusion analysis for behavioral and stochastic factors in pedestrian trajectory prediction.

Table 3. Ablation results for long-term cases on SDD.

Tricks	ADE	FDE
None (LSN-GTDA)	55.25	68.27
TMSS (LSN-GTDA)	49.54	65.35
PNMSS (LSN-GTDA)	48.42	64.24
TMSS + PNMSS (LSN-GTDA)	**31.91**	**41.45**

Additionally, Figure 7 indicates the impact of diversity brought from both multimodalities of various choices of M_s and M_b; the results show consistent ADE improvements for various M_b while increasing M_s, which indicates the effectiveness of multimodality. According to Table 3, the baseline of "None (LSN-GTDA)" lacking the multimodality strategy performs much worse than other baselines using at least one multimodality strategy. Through the multimodality decomposition of the path and the target, the stochastic factor and the behavioral factor of the uncertainty correspond. Thus, the degree of uncertainty is reduced, and the prediction error is also reduced, which highlights the importance of factorizing target and path multimodality for diverse and future trajectory modeling.

4.6. Parameter Workload Analysis

In addition, Table 4 shows the inference time and parameter workload of state-of-the-art deep learning-based models. Our LSN-GTDA method is lighter than other methods, only slightly higher than Y-Net by about 13.4% since LSN-GTDA has a symmetrical network. However, we observe that in the training stage, the proposed LSN-GTDA algorithm converges after 50 epochs of iteration, while the Y-Net baseline needs 200 epochs to converge. Therefore, compared with the Y-Net baseline, our LSN-GTDA algorithm shortened the training convergence time by 75% while increasing the number of network model parameters by only 13.4%. As shown in Table 4, the inference time of our method shows low-latency trajectory prediction in the training stage, which indicates that the proposed LSN-GTDA model has realized the real-time forecasting goal by using the proposed complete response mechanism-based thermal diffusion process rather than the traditional diffusion models based on time-consuming denoising processes. In the test stage, LSN-GTDA costs 23.86 s for each long-term scene in SDD, and also obtained better results than Y-Net, which shows the superiority of our method.

Table 4. Comparisons of inference times at different batch-size settings (from 1 to 1000) and the parameter size of different Pytorch models on one NVIDIA TITAN Xp GPU with 12 GB memory.

Model	Inference Time @Batchsize (ms)					Parameters
	1	50	100	500	1000	
SGNet [19]	76	90	125	213	591	7.27 M
MemoNet [62]	48	63	78	146	368	5.04 M
Eqmotion [19]	31	39	54	102	189	2.89 M
Y-Net [18]	15	20	27	51	87	1.57 M
LSN-GTDA(Ours)	17	23	28	55	94	1.78 M

4.7. Conclusions and Discussion

In this paper, we propose a human trajectory forecasting method by a novel mechanism of global energy analysis, which models the uncertainty of prediction from the behavioral factor and the stochastic factor by the zero-input response and the zero-state response. Our method reveals the inherent principle of the multiple modalities of the human trajectory and provides a solution to learn the principle of human motion from UAV data.

For downstream applications, multimodal human trajectory prediction can be applied to a wide range of fields involving human–robot interaction. By incorporating uncertainty modeling, accurate and reliable future trajectories can be generated, which is crucial for supporting autonomous driving decisions. Furthermore, the proposed LSN-GTDA method offers interpretability in addressing uncertainty, making it a promising solution for more dynamic and interactive domains, such as human–computer interaction, action recognition, UAV-based object tracking, and person re-identification.

Expanding human trajectory prediction from UAV scenarios to pedestrian tracking is feasible and can be beneficial. They share similar dynamics of human movement, and the sensor fusion may assist data from UAVs to enhance tracking models. However, there are also some challenges such as their environmental differences; UAVs usually operate in open areas while pedestrians often encounter obstacles, varying terrain, and a higher density of objects, which may lead to a loss or mismatch in general tracking algorithms, e.g., SORT and ByteTrack. One feasible solution is to design a person re-identification algorithm under a broad UAV scenario to match the pedestrians detected in the next frame after the tracking target is lost; this may quickly recover the tracking.

In the future, we will expand the proposed method to a scenario of several pedestrians such as a group and try to promote it with related work on person and group re-identification [63,64] under UAV scenarios.

Author Contributions: Conceptualization L.M. and L.Z.; methodology, L.M., B.W., M.F. and L.Z.; software, L.M., B.W., L.Z. and M.F.; validation, L.M. and M.F.; formal analysis, L.M. and M.Y.; investigation, L.M., M.F. and Y.Z.; resources, L.M., L.Z. and L.J.; data curation, L.M., M.F., Y.Z., M.Y. and L.J.; writing—original draft preparation, L.M., B.W. and M.F.; writing—review and editing, L.M. and L.Z.; visualization, L.M., M.Y. and M.F.; supervision, L.M. and L.Z.; project administration, L.Z.; funding acquisition, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62306218, the Nature Science Foundation of Hubei Province of China under Grant 2023AFB070, the National Natural Science Foundation of China under Grant 62203177, and the Department of Science and Technology of Hubei Province of China under Grant 2022EHB015.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the first author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, S.; Cao, L.; Li, Y. Lightweight pedestrian detection network for UAV remote sensing images based on strideless pooling. *Remote. Sens.* **2024**, *16*, 2331. [CrossRef]
2. Zhu, Y.; Wang, T.; Zhu, S. Adaptive multi-pedestrian tracking by multi-sensor: Track-to-track fusion using monocular 3D detection and MMW radar. *Remote. Sens.* **2022**, *14*, 1837. [CrossRef]
3. Gómez Arnaldo, C.; Zamarreño Suárez, M.; Pérez Moreno, F.; Delgado-Aguilera Jurado, R. Path Planning for Unmanned Aerial Vehicles in Complex Environments. *Drones* **2024**, *8*, 288. [CrossRef]
4. Cui, X.; Wang, C.; Xiong, Y.; Mei, L.; Wu, S. More Quickly-RRT*: Improved Quick Rapidly-exploring Random Tree Star algorithm based on optimized sampling point with better initial solution and convergence rate. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108246. [CrossRef]
5. Zhang, G.; Liu, T.; Ye, Z. Dynamic Screening Strategy Based on Feature Graphs for UAV Object and Group Re-Identification. *Remote. Sens.* **2024**, *16*, 775. [CrossRef]
6. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium, Las Vegas, NE, USA, 19–22 July 2020; pp. 1929–1934.
7. Mei, L.; He, Y.; Fishani, F.J.; Yu, Y.; Zhang, L.; Rhodin, H. Learning Domain-Adaptive Landmark Detection-Based Self-Supervised Video Synchronization for Remote Sensing Panorama. *Remote. Sens.* **2023**, *15*, 953. [CrossRef]
8. Liu, Y.; Liao, Y.; Lin, C.; Jia, Y.; Li, Z.; Yang, X. Object tracking in satellite videos based on correlation filter with multi-feature fusion and motion trajectory compensation. *Remote. Sens.* **2022**, *14*, 777. [CrossRef]
9. Zhang, S.; Li, Y.; Wu, X.; Chu, Z.; Li, L. MRG-T: Mask-Relation-Guided Transformer for Remote Vision-Based Pedestrian Attribute Recognition in Aerial Imagery. *Remote. Sens.* **2024**, *16*, 1216. [CrossRef]
10. Bennewitz, M.; Burgard, W.; Thrun, S. Learning motion patterns of persons for mobile service robots. In Proceedings of the IEEE International Conference on Robotics and Automation, Washington, DC, USA, 11–15 May 2002; Volume 4, pp. 3601–3606.
11. Thrun, S. Probabilistic robotics. *Commun. ACM* **2002**, *45*, 52–57. [CrossRef]
12. Li, K.; Guo, D.; Chen, G.; Liu, F.; Wang, M. Data Augmentation for Human Behavior Analysis in Multi-Person Conversations. In Proceedings of the ACM International Conference on Multimedia, Ottawa, ON, Canada, 28 October–3 November 2023; pp. 9516–9520.
13. Mei, L.; Yu, M.; Jia, L.; Fu, M. Crowd Density Estimation via Global Crowd Collectiveness Metric. *Drones* **2024**, *8*, 616. [CrossRef]
14. Mei, L.; Lai, J.; Feng, Z.; Chen, Z.; Xie, X. Person re-identification using group constraint. In Proceedings of the Intelligence Science and Big Data Engineering, Visual Data Engineering: 9th International Conference, IScIDE 2019, Nanjing, China, 17–20 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 459–471.
15. Takumi, K.; Watanabe, K.; Ha, Q.; Tejero-De-Pablos, A.; Ushiku, Y.; Harada, T. Multispectral object detection for autonomous vehicles. In Proceedings of the Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 35–43.
16. Mei, L.; Lai, J.; Chen, Z.; Xie, X. Measuring crowd collectiveness via global motion correlation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October 2019; pp. 1222–1231.
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

18. Mangalam, K.; An, Y.; Girase, H.; Malik, J. From goals, waypoints & paths to long term human trajectory forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 15233–15242.
19. Wang, C.; Wang, Y.; Xu, M.; Crandall, D.J. Stepwise goal-driven networks for trajectory prediction. *IEEE Robot. Autom. Lett.* **2022**, *7*, 2716–2723. [CrossRef]
20. Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.H.; Adeli, E.; Malik, J.; Gaidon, A. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 759–776.
21. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2016; pp. 961–971.
22. Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; Alahi, A. Social gan: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2255–2264.
23. Liang, J.; Jiang, L.; Niebles, J.C.; Hauptmann, A.G.; Fei-Fei, L. Peeking into the future: Predicting future person activities and locations in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5725–5734.
24. Helbing, D.; Molnar, P. Social force model for pedestrian dynamics. *Phys. Rev. E* **1995**, *51*, 4282. [CrossRef]
25. Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; Savarese, S. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1349–1358.
26. Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.Z.; Vo, M.; Malik, J. Long-term human motion prediction with scene context. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 387–404.
27. Liang, J.; Jiang, L.; Murphy, K.; Yu, T.; Hauptmann, A. The garden of forking paths: Towards multi-future trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10508–10518.
28. Tomasello, M.; Carpenter, M.; Call, J.; Behne, T.; Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.* **2005**, *28*, 675–691. [CrossRef]
29. Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murgesan, K.; Mattei, N.; Rossi, F.; et al. Thinking fast and slow in AI. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 15042–15046.
30. Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; Savarese, S. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–10.
31. Lee, N.; Choi, W.; Vernaza, P.; Choy, C.B.; Torr, P.H.; Chandraker, M. Desire: Distant future prediction in dynamic scenes with interacting agents. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 336–345.
32. Salzmann, T.; Ivanovic, B.; Chakravarty, P.; Pavone, M. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In Proceedings of the European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 683–700.
33. Maeda, T.; Ukita, N. Fast inference and update of probabilistic density estimation on trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9795–9805.
34. Zhou, B.; Tang, X.; Wang, X. Measuring crowd collectiveness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Sydney, Australia, 1–8 December 2013; pp. 3049–3056.
35. Yue, J.; Manocha, D.; Wang, H. Human trajectory prediction via neural social physics. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 376–394.
36. Wong, C.; Xia, B.; Hong, Z.; Peng, Q.; Yuan, W.; Cao, Q.; Yang, Y.; You, X. View Vertically: A hierarchical network for trajectory prediction via fourier spectrums. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 682–700.
37. Lin, W.; Mi, Y.; Wang, W.; Wu, J.; Wang, J.; Mei, T. A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Trans. Image Process.* **2016**, *25*, 1674–1687. [CrossRef] [PubMed]
38. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 14347–14356.

39. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11461–11471.
40. Xie, H.; Yang, Z.; Zhu, H.; Wang, Z. Striking a balance: Unsupervised cross-domain crowd counting via knowledge diffusion. In Proceedings of the ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 6520–6529.
41. Shi, L.; Wang, L.; Zhou, S.; Hua, G. Trajectory unified transformer for pedestrian trajectory prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 9675–9684.
42. Wong, C.; Xia, B.; Zou, Z.; Wang, Y.; You, X. SocialCircle: Learning the Angle-based Social Interaction Representation for Pedestrian Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 19005–19015.
43. Kim, S.; Chi, H.g.; Lim, H.; Ramani, K.; Kim, J.; Kim, S. Higher-order Relational Reasoning for Pedestrian Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 15251–15260.
44. Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; Litany, O. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13756–13766.
45. Mao, W.; Xu, C.; Zhu, Q.; Chen, S.; Wang, Y. Leapfrog diffusion model for stochastic trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 5517–5526.
46. Mei, L.; Lai, J.; Xie, X.; Zhu, J.; Chen, J. Illumination-invariance optical flow estimation using weighted regularization transform. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 495–508. [CrossRef]
47. Mei, L.; Chen, Z.; Lai, J. Geodesic-based probability propagation for efficient optical flow. *Electron. Lett.* **2018**, *54*, 758–760. [CrossRef]
48. Hs, C.; Jaeger, J. *Conduction of Heat in Solids*; Oxford University Press: Oxford, UK, 1959.
49. Oppenheim, A.V.; Willsky, A.S.; Nawab, S.H.; Ding, J.J. *Signals and Systems*; Prentice Hall: Upper Saddle River, NJ, USA, 1997; Volume 2.
50. Goodfellow, I.; Bengio, Y.; Courville, A. Softmax units for multinoulli output distributions. In *Deep Learning*; MIT Press: Cambridge, MA, USA, 2018.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NE, USA, 27–30 June 2016; pp. 770–778.
52. Sadeghian, A.; Kosaraju, V.; Gupta, A.; Savarese, S.; Alahi, A. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv* **2018**, arXiv:1805.07663.
53. Pellegrini, S.; Ess, A.; Schindler, K.; Van Gool, L. You'll never walk alone: Modeling social behavior for multi-target tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 261–268.
54. Lerner, A.; Chrysanthou, Y.; Lischinski, D. Crowds by example. In *Proceedings of the Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2007; Volume 26, pp. 655–664.
55. Alahi, A.; Ramanathan, V.; Fei-Fei, L. Socially-aware large-scale crowd forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2203–2210.
56. Bae, I.; Oh, J.; Jeon, H.G. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Pairs, France, 2–3 October 2023; pp. 10017–10029.
57. Bhattacharyya, A.; Hanselmann, M.; Fritz, M.; Schiele, B.; Straehle, C.N. Conditional Flow Variational Autoencoders for Structured Sequence Prediction. In Proceedings of the 4th workshop on Bayesian Deep Learning of NeurIPS 2019, Vancouver, BC, Canada, 13 December 2019; pp. 1–11.
58. Deo, N.; Trivedi, M.M. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv* **2020**, arXiv:2001.00735.
59. Liang, J.; Jiang, L.; Hauptmann, A. Simaug: Learning robust representations from simulation for trajectory prediction. In Proceedings of the European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 275–292.
60. Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; Lu, J. Stochastic trajectory prediction via motion indeterminacy diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 17113–17122.
61. Feng, A.; Han, C.; Gong, J.; Yi, Y.; Qiu, R.; Cheng, Y. Multi-Scale Learnable Gabor Transform for Pedestrian Trajectory Prediction From Different Perspectives. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 13253–13263. [CrossRef]

62. Xu, C.; Mao, W.; Zhang, W.; Chen, S. Remember intentions: Retrospective-memory-based trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6488–6497.
63. Mei, L.; Lai, J.; Feng, Z.; Xie, X. From pedestrian to group retrieval via siamese network and correlation. *Neurocomputing* **2020**, *412*, 447–460. [CrossRef]
64. Mei, L.; Lai, J.; Feng, Z.; Xie, X. Open-world group retrieval with ambiguity removal: A benchmark. In Proceedings of the IEEE International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 584–591.