DiffUFlow: Robust Fine-grained Urban Flow Inference with Denoising Diffusion Model

Yuhao Zheng
Central South University
Changsha, China
8212200810@csu.edu.cn

Lian Zhong
Central South University
Changsha, China
8203201009@csu.edu.cn

Senzhang Wang*
Central South University
Changsha, China
szwang@csu.edu.cn

Yu Yang
The Hong Kong Polytechnic University
Hong Kong, China
cs-yu.yang@polyu.edeu.hk

Weixi Gu
China Academy of Industrial Internet
Beijing, China
guweixigavin@gmail.com

Junbo Zhang
JD Intelligent Cities Research
Beijing, China
msjunbozhang@outlook.com

Jianxin Wang
Central South University
Changsha, China
jxwang@csu.edu.cn

ABSTRACT

Inferring the fine-grained urban flows based on the coarse-grained flow observations is practically important to many smart city-related applications. However, the collected urban flows are usually rather unreliable, may contain noise and sometimes are incomplete, thus posing great challenges to existing approaches. In this paper, we present a pioneering study on robust fine-grained urban flow inference with noisy and incomplete urban flow observations, and propose a denoising diffusion model named DiffUFlow to effectively address it with an improved reverse diffusion strategy. Specifically, a spatial-temporal feature extraction network called STFormer and a semantic features extraction network called ELFetcher are proposed. Then, we overlay the extracted spatial-temporal feature map onto the coarse-grained flow map, serving as a conditional guidance for the reverse diffusion process. We further integrate the semantic features extracted by ELFetcher to cross-attention layers, enabling the comprehensive consideration of semantic information for fine-grained flow inference. Extensive experiments on two large real-world datasets validate the effectiveness of our method compared with the state-of-the-art baselines.

CCS CONCEPTS

• **Information systems** → **Location based services**.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614842

KEYWORDS

Spatial-temporal data mining, Urban flow inference, Denoising diffusion model

ACM Reference Format:
Yuhao Zheng, Lian Zhong, Senzhang Wang[1], Yu Yang, Weixi Gu, Junbo Zhang, and Jianxin Wang. 2023. DiffUFlow: Robust Fine-grained Urban Flow Inference with Denoising Diffusion Model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3583780.3614842

1 INTRODUCTION

Accurately depicting the fine-grained patterns of urban flows, such as taxi flows, bicycle flows, and pedestrian trajectories, is crucial for various smart city applications, including urban planning, urban renewal, and traffic management [1]. However, acquiring these observational insights requires the deployment of numerous sensors in different regions of the city, resulting in significant daily operational and maintenance costs. For instance, a substantial number of cameras need to be installed at the intersections of the traffic network to perceive the overall traffic conditions of the entire city. However, in reality, usually only a small number of sensors are deployed for collecting urban flow data due to the limited budget. Consequently, inferring fine-grained urban traffic flows based on coarse-grained sensor observations has emerged as a pivotal research topic, and attracted rising research attention in recent years [2, 3].

Motivated by the super-resolution methods in computer vision (CV), methods like SRCNN [4], VDSR [5], and SRResNet [6] have been proposed to address the spatial-temporal data super-resolution task and have achieved superior performance than traditional methods [7]. However, directly borrowing the super-resolution methods from the CV area will fall short in ignoring the complex spatial-temporal correlations of the traffic data and some external factors

*Corresponding author.

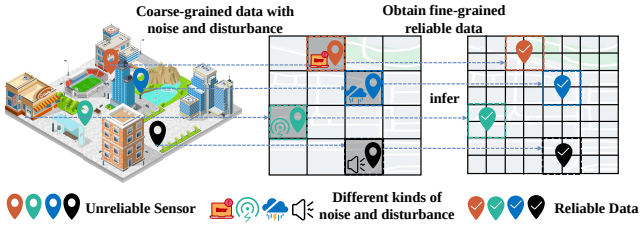

Figure 1: Illustration of inferring fine-grained urban flows with unreliable data. The grey cell regions denote the unreliable urban flows with noise and missing values.

such as weather, temperature, and holidays. To address this limitation, methods like UrbanFM [1], UrbanPy [8], and DeepLGR [9] have been proposed, which incorporate the specific characteristics of urban traffic flow data into the model design and achieve significant performance improvement. To address the issue that the coarse-grained urban flows collected in real-world scenarios can be sparse and incomplete, MT-CSR [10] establishes a joint network for both urban flow data imputation and super-resolution. It effectively accomplishes both the data completion and super-resolution tasks while considering the complex spatial-temporal correlations and external semantic features. However, a major limitation of MT-CSR is that it still assumes that the observed urban flows are reliable without considering the effect of noise.

As shown in Figure 1, in practical scenarios, the collected urban flow data are usually rather unreliable, which are sparse and contain noise due to the unreliability of road sensors and data transmission/storage error. To address this issue, this paper for the first time infers the fine-grained urban traffic flows based on the unreliable coarse-grained urban flow observations. Compared with previous works, the problem addressed in this study is more challenging due to the following reasons.

- **The complex and dynamic spatial-temporal correlations.** Different from images, the spatial-temporal dependencies of the urban flow data are more complex and dynamic, and thus are hard to capture by existing super-resolution model architectures [11]. Existing methods mainly adopt CNN, GNN, and LSTM to capture the spatial-temporal correlations of the urban flows [12, 13], which are locally biased and can only model the short-term temporal dependency. Thus such methods cannot effectively capture the long-term temporal dependencies, such as periodicity and trend. GNN-based models [14, 15] try to capture the static spatial dependencies, which are not effective to learn the dynamic urban flow interaction patterns. Hence, a new method is required to more effectively capture the complex temporal correlations and the dynamic spatial dependencies.
- **Robust urban flow super-resolution with unreliable data.** Existing fine-grained urban flow inference methods generally assume the observed urban flows are credible. However, as we argued before, the data are actually noisy and incomplete due to the devise errors and limited sensor coverage range, making existing methods not directly applicable. To achieve a robust urban flow super-resolution result, we need to first eliminate the negative effect of the noisy and sparse data. How to design a new framework that can effectively integrate the spatial-temporal correlations of the urban flows and achieve a robust reference result is also challenging.

To address the above challenges, this paper proposes a Diffusion-based robust fine-grained Urban Flow inference model (DiffUFlow for short). As the first generative approach for fine-grained urban flow inference, DiffUFlow is designed as a conditional denoising diffusion probabilistic model (DDPM) with U-Net [16] as the backbone architecture. Specifically, to more effectively capture the complex and dynamic spatial-temporal correlations, we design a transformer based spatial-temporal feature extraction network named STFormer. A semantic feature extraction network named ELFetcher is also proposed to extract features containing external factors and land features. Next, to fully consider the unique characteristics of urban flow data during the diffusion process, we flexibly integrate these two types of features into the U-Net and propose a reverse diffusion strategy in the conditional DDPM. Specifically, in each denoising step, we expand the input channels of the reverse denoising process by overlaying the extracted feature maps and the coarse-grained flow map onto the flow map input. Note that both the extracted features and the coarse-grained flows can be considered as the conditional guidance for the reverse diffusion process. U-Net, STFormer, and ELFetcher are jointly trained in an end-to-end manner in each denoising step. Our main contributions are summarized as follows.

- We for the first time study the fine-grained urban flow inference problem with unreliable data, and propose a denoising diffusion-based generative model DiffUFlow to effectively address it.
- A transformer-based model is proposed to effectively extract the complex and dynamic spatial-temporal features which serve as a conditional guidance of the reverse denoising process.
- Extensive experiments are conducted on two large real-world datasets. Experimental results demonstrate the superiority of DiffUFlow by comparison with existing state-of-the-art approaches.

The remainder of the paper is organized as follows. We will first briefly review related work in Section 2. Notations and problem definition will be introduced in Section 3. DiffUFlow model will be introduced in Section 4, followed by experimental results in Section 5. Finally, we will conclude the paper in Section 6.

2 RELATED WORK

2.1 Urban Flows Representation Learning

Recently, various deep learning models, including convolutional neural network (CNN), recurrent neural network (RNN), and graph neural network (GNN), are widely adopted to learn the spatial-temporal representations of the urban flow data for fine-grained urban flow inference [17, 18]. ST-ResNet [19] first employs a CNN with residual connections to capture the spatial correlations and then construct several branches of historical data based on temporal semantics to extract the temporal correlations. Similar idea is

adopted by subsequent works [20] to learn the urban flow representation. RNN-based methods, such as GRU [21] and LSTM[22], are introduced to extract the spatial-temporal correlations, by constructing a sequential architecture. GNN models [23, 24] are also introduced into the urban flow inference task, due to its superiority in modeling graph structural data. Different from existing methods, DiffUFlow comprehensively considers the dynamic and long-range spatial-temporal dependencies and designs a tranfomer-based module to learn the urban flow representation.

2.2 Spatial-Temporal Data Super-Resolution

In recent years, image super-resolution methods have been applied to spatial-temporal data super-resolution tasks. For example, methods like SRCNN [4], VDSR [5], and SRResNet [6] have achieved better performance than traditional methods [7]. However, due to the complex spatial-temporal correlations and the effect of external factors, there are fundamental differences between image super-resolution and spatial-temporal data super-resolution. UrbanFM [1] devises an external factor fusion network to extract and fuse external features such as weather, temperature, and holidays with the inference network. UrbanPy [8] further enhances the performance of UrbanFM at high upscaling rates. DeepLGR [9] introduces a local feature extraction module for nearby information, a global context module for expanding the field of view, and a region-specific predictor for reducing the number of network parameters. Considering that the coarse-grained urban flows collected in real-world scenarios are not always complete, Li et al. [10] proposes a multi-task learning model named MT-CSR that can conduct data completion and super-resolution simultaneously. However, existing works generally assume that the coarse-grained urban flows are reliable, but ignore the negative effect of noise and data sparsity issues.

2.3 Denoising Diffusion Probabilistic Model

Diffusion models have emerged as the state-of-the-art deep generative models. Besides its superior performance on image synthesis [25], diffusion model has shown great potential in various domains, including computer vision [26], natural language processing [27], temporal data modeling [28, 29], multi-modal modeling [30, 31], robust learning [32], molecular graph modeling [33], material design [34] and inverse problem solving [35]. Motivated by the great success of diffusion models, this paper for the first time designs a conditional denoising diffusion model for fine-grained urban flow inference.

3 NOTATIONS AND PROBLEM DEFINITION

We will first give some definitions to help us state the studied problem, and then present a formal problem definition.

Definition 1. ***Cell region.*** *We divide a city into a grid map consisting of IxJ cell regions based on latitude and longitude. We denote all the cell regions as* $R = \{r_{1,1}, ...r_{i,j}, ...r_{I,J}\}$*, where* $r_{i,j}$ *is the i-th row and j-th column cell region of the grid map.*

Definition 2. ***Urban flow map*** *Let* $\mathcal{T}$ *represent a collection of urban flow trajectories. When considering a specific cell region* $r_{i,j}$*, the associated inflow and outflow map of urban flows during the time slot t can be defined as follows:*

$$X^t_{in,i,j} = \sum_{f^t \in \mathcal{T}} \{f^{t-1} \notin r_{i,j} \cap f^t \in r_{i,j}\}$$

$$X^t_{out,i,j} = \sum_{f^t \in \mathcal{T}} \{f^t \in r_{i,j} \cap f^{t+1} \notin r_{i,j}\}$$

where $f \in \mathcal{T}$ *represents a trajectory within the collection.* $f^{t-1} \notin r_{i,j}$ *indicates that the trajectory* f *at time t-1 does not fall into the region* $r_{i,j}$*. Conversely,* $f^t \in r_{i,j}$ *denotes that the trajectory* f *at time t falls into the region* $r_{i,j}$*.* $\cap$ *represents the intersection operator. To represent the inflows and outflows of all regions at time t, we introduce the urban flow tensor* $x^t \in \mathbb{R}^{2 \times I \times J}$*.*

Definition 3. ***Coarse- and fine-grained urban flow spatial maps.*** *A coarse-grained urban flow spatial map represents the observed urban flows derived from the flow sensors. It is generated by integrating neighboring grids within an* $N \times N$ *range from a fine-grained urban flow map, where N is the upscaling factor. We denote the coarse-grained and fine-grained urban flow maps at time t as* $x^t_{cg} \in \mathbb{R}^{2 \times I \times J}$ *and* $x^t_{fg} \in \mathbb{R}^{2 \times NI \times NJ}$*, respectively. Note that coarse-grained spatial maps may contain noise. We represent the coarse-grained spatial map with noise data as* $x^t_{cg,no} \in \mathbb{R}^{2 \times I \times J}$*.*

Definition 4. ***External factor vector.*** *External factors include temperature, windspeed, holiday, e.t. We represent these external factors in a time slot t as a vector* $\mathcal{E}_t \in \mathbb{R}^l$*, where l is the feature length.*

Definition 5. ***Land feature matrix.*** *We collect POI category distribution and the structural attributes of road networks in each cell region as its land features, represented as a matrix* $\mathcal{P} \in \mathbb{R}^{K \times I \times J}$*, where K is the feature category.*

Problem Statement. Given an upscaling factor N, a set of historical observations of urban flows $X_{history}$, coarse-grained urban flow spatial map with noisy and incomplete data $x^t_{cg,no} \in \mathbb{R}^{2 \times I \times J}$, the external factors $\mathcal{E}^t$, and the land features $\mathcal{P}$, our target is to infer the fine-grained urban flow map $x^t_{fg} \in \mathbb{R}^{2 \times NI \times NJ}$.

输入输出

4 METHODOLOGY

Figure 2 shows the framework of the proposed DiffUFlow, which consists of three major parts: conditional denoising diffusion probabilistic model (DDPM), spatial-temporal feature extraction network STFormer and semantic features extraction network ELFetcher. We propose a conditional DDPM with U-Net as its backbone as a generative approach for fine-grained urban flow inference. The model contains noise addition and denoising processes, which offers unique advantages when dealing with data containing noise and disturbance, and thus can provide robust reference result. To fully consider the external features and semantic features, we propose to expand the input channels and employ a cross-attention mechanism to integrate these features as conditional guidance during the reverse denoising process. We will elaborate this part in Sections 4.1. STFormer aims to extract spatial-temporal features and ELFecher is designed for capturing semantic features. STFormer contains Pre-Conv Block and follows the Vision Transformer architecture, enabling the learning of both local and global spatial correlations. This part will be introduced in Section 4.2. ELFecher extracts the

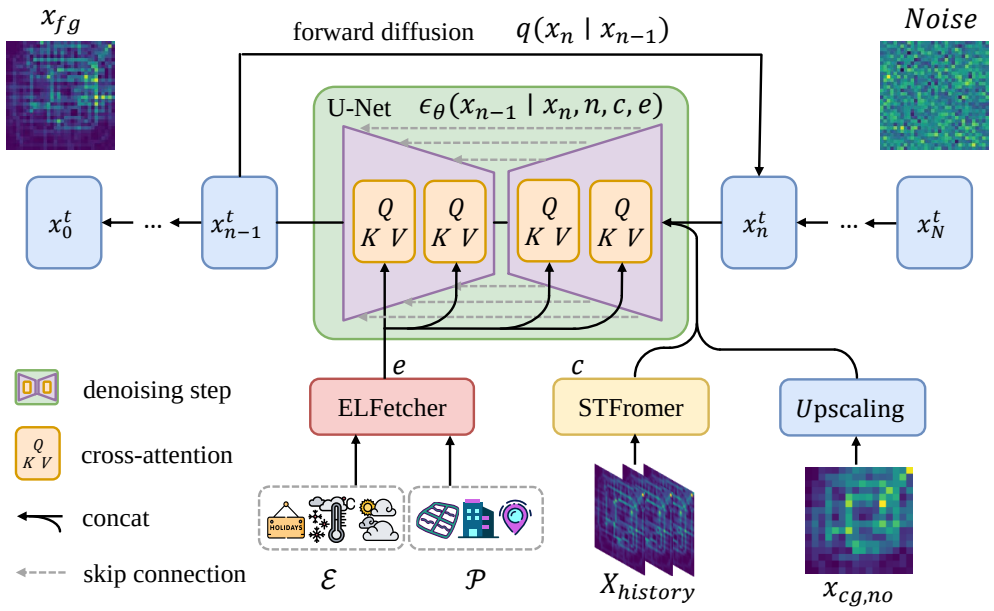

Figure 2: Framework of the proposed DiffUFlow model

semantic features by first adopting two decoders separately, and then performing adaptive fusion, which will be introduced in detail in Section 4.3.

4.1 Conditional Denoising Diffusion Probabilistic Model

As a class of generative models, Denoising Diffusion Probabilistic Models [36] have demonstrated state-of-the-art performance on various data modalities. Diffusion models aim to learn a mapping from the latent space to the signal space by sequentially learning to remove noise in a reverse process that is added sequentially in a Markovian fashion during a so-called forward process. These two processes, therefore, form the basic framework of diffusion model. For simplicity, we will first introduce the unconditional case, and then discuss the modifications for the conditional case.

The forward process is parameterized as follows,

$$q(x_1, \ldots, x_N|x_0) = \prod_{n=1}^{N} q(x_n|x_{n-1}) \tag{1}$$

$$q\left(x_n \mid x_{n-1}\right) = \mathcal{N}\left(x_n; \sqrt{1-\beta_n}x_{n-1}, \beta_n \mathbf{I}\right), \tag{2}$$

where $q\left(x_n \mid x_{n-1}\right)$ denotes the conditional Gaussian distribution and the forward-process variances β_n adjust the noise level. Equivalently, x_n can be expressed in the closed form as follows,

$$x_n = \sqrt{\alpha_n}x_0 + (1-\alpha_n)\,\epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0,1)$ and $\alpha_n = \sum_{i=1}^{n}(1-\beta_n)$.

The backward process is parameterized as

$$p_\theta(x_0, \ldots, x_{n-1}|x_N) = p(x_N)\prod_{t=1}^{N} p_\theta(x_{n-1}|x_n), \tag{4}$$

where $x_N \sim \mathcal{N}(0,1)$. Again, $p_\theta(x_{n-1}|x_n)$ is assumed to follow normal-distributed (with diagonal covariance matrix) with learnable parameters. Using a particular parameterization of $p_\theta(x_{n-1}|x_n)$, the reverse process can be trained through optimizing the following objective,

$$L = \min_\theta \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1), n\sim\mathcal{U}(1,N)}||\epsilon - \epsilon_\theta(x_n, n)||_2^2, \tag{5}$$

where $\mathcal{D}$ refers to the data distribution and $\epsilon_\theta(x_n, n)$ is parameterized using a neural network. This objective can be seen as a weighted variational bound on the negative log-likelihood that down-weights the importance of terms at small n, i.e., at small noise levels.

The above description has focused on the unconditional diffusion process. However, in the scenario of urban flow inference, we need to adopt a conditional variant for fully considering the unique features of urban flow data. Therefore, we devise the reverse denoising process by incorporating additional information as embedded conditions.

The original distribution x_0 of the diffusion process is fine-grained urban flow map x_{fg}^t, which transforms into an isotropic Gaussian distribution $\mathcal{N}(0,1)$ by forward diffusion in N steps. We attempt to recover the original distribution by considering the rich features of urban flows extracted by STFormer and ELFecher networks. STFormer extracts spatial-temporal features c^t from historical data $X_{history}^t$, while ELFecher extracts and fuses features e^t from external factors $\mathcal{E}^t$ and the land features $\mathcal{P}$. Detailed description on these two networks will be provided in Sections 4.2 and 4.3.

Next, we employ two different embedding strategies for the extracted features c^t and e^t. During each reverse denoising step n, we expand the input channels of the traditional reverse diffusion

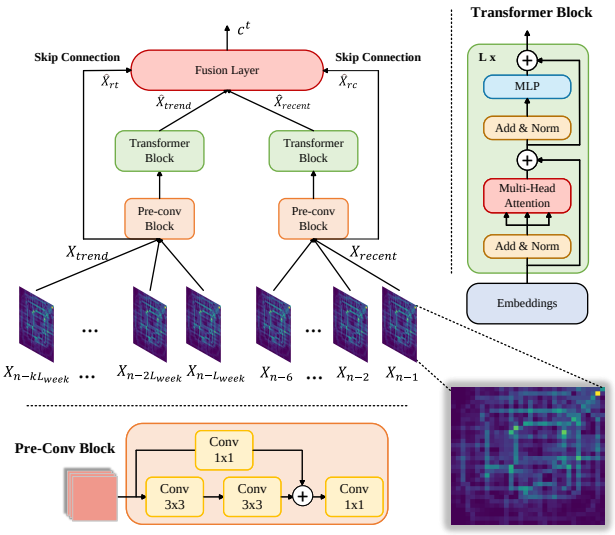

Figure 3: Framework of STFormer

process. Specifically, we overlay the spatial-temporal feature map c^t extracted by STFormer and the coarse-grained flow map $x^t_{cg,no}$ onto the flow map of this step, serving as two conditional guidance for the reverse diffusion process. Additionally, the semantic features e^t are mapped to the intermediate layers of U-Net through the cross-attention operation, which facilitates the fusion of external factors and land features at each step n.

The final loss function of denoising step n for DiffUFlow is as follows,

$$\begin{aligned} L &= \min_{\theta} \mathbb{E}\|\epsilon - \epsilon_\theta(concat(x^t_n, n, c^t), e^t)\|_2^2, \\ c^t &= f_\theta(X^t_{history}), \\ e^t &= g_\theta(\mathcal{E}^t, \mathcal{P}) \end{aligned} \tag{6}$$

where f_θ and g_θ denote STFormer and ELFecher, respectively. Note that ϵ_θ, f_θ, and g_θ are jointly optimized.

4.2 STFormer

To fully capture the complex and dynamic spatial-temporal features of urban flows, we exploit the trend data and the recent data when calculating the fine-grained flow map. We introduce a spatial-temporal feature extraction network called STFormer, which effectively captures intricate and dynamic spatial-temporal correlations. STFormer comprises Pre-Conv Block and Vision Transformer [37], enabling the learning of both local and global spatial correlations. In addition, the historical data is directly connected to the output through skip connections, allowing for the full use of the historical data as the base information.

Specifically, we use a symmetric structure to process long-term temporal data X_{trend} and short-term temporal data X_{close}. The input data X_{trend} and X_{close} are first pre-processed by a transform block consisting of a convolutional network Pre-conv Block, which focuses on capturing local correlations in the adjacent regions, while the global features are captured by the Vision Transformer (ViT). After ViT extracts the temporal features $\hat{X}_{trend}$ and $\hat{X}_{close}$, the residual components $\hat{X}_{rt}$ and $\hat{X}_{rc}$ are simultaneously retrieved using a skip connection. Finally, STFormer fuses the four components $(\hat{X}_{trend}, \hat{X}_{close}, \hat{X}_{rt}, \hat{X}_{rc})$ in the fusion block to produce the final temporal feature map c^t as follows,

$$c^t = w_c \cdot \hat{X}_{close} + w_t \cdot \hat{X}_{trend} + w_{rc} \cdot \hat{X}_{rc} + w_{rt} \cdot \hat{X}_{rt}, \tag{7}$$

where $\cdot$ represents element-wise multiplication, and w is the learnable weight parameter of each component.

4.3 ELFetcher

Urban flows can be significantly affected by various external factors and land features. To further extract such features, we design a ELFetcher module to learn and integrate external features and land features, and incorporate them into the reverse diffusion process as conditional guidance.

The external factors contain both continuous factors and categorical factors. Continuous factors are directly concatenated and categorical factors are input into separate embedding layers. Next, we employ an external factor encoder and a land feature encoder to extract their features separately. Specifically, the external factor encoder consists of two MLP modules with Swish activation functions. ResNet-18 [38] is used as the land feature encoder. Then, linear projection and normalization are applied, and the resulting feature vectors are concatenated. Finally, a 2-layer MLP is used to the concatenated features, yielding the final feature vector $\mathbf{e}^t$.

To flexibly integrate the feature vectors into the reverse diffusion process as a condition, we employ a cross-attention mechanism to map the feature vector $\mathbf{e}^t$ to the intermediate representation layer of the U-Net as follows,

$$\begin{aligned} MultiHead &= Concat(head_1, \ldots, head_i) * W \\ head_i &= Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \\ Q &= W_Q^{(i)} \cdot \varphi_i(x_t), K = W_K^{(i)} \cdot e^t, V = W_V^{(i)} \cdot e^t. \end{aligned} \tag{8}$$

where W is the parameter matrix, $e^t = g_\theta(\mathcal{E}^t, \mathcal{P})$ and $\varphi_i(x_n) \in \mathbb{R}^{N \times d_\epsilon^i}$ denote a (flattened) intermediate representation of the U-Net implementing ϵ_θ.

5 EXPERIMENT

5.1 Experiment Setup 数据集

5.1.1 Dataset. We evaluate the performance of the proposed model on two taxi trajectory datasets, BJTaxi and NYCTaxi. BJTaxi contains the taxi trips in Beijing covering from March 1 to June 30 in 2015. We partition the entire data into non-overlapping training, testing and validation sets by a ratio of $\{7:2:1\}$. NYCTaxi contains more than 160 million records of taxi trips in New York City, spanning from January 1 to December 31 in 2015. Each trip record contains information such as the date and time of pick-up and drop-off, pick-up and drop-off locations, trip distance, itemized fare, rate type, payment type, and the number of passengers reported by the driver. To utilize this dataset, we use the first 11 months' data for training and validation, while the data of the last month is used for testing. The dataset descriptions are shown in Table 1.

Table 1: Dataset Description

Dataset	BJTaxi	NYCTaxi
Latitude	/	(40.71,40.765)
longitude	/	(-74.01,-73.972)
Time span	3/1/2015-6/30/2015	1/1/2015-12/31/2015
Time interval	30 minutes	1 hour
Size	5760	8760
Upscaling	$16 \times 16 \longrightarrow 32 \times 32$	

5.1.2 Data Preprocessing. To test the robustness of our model with noisy and incomplete urban flow data, we conduct three data degradation operations: data deformation, data missing, and adding noise. By combining and overlaying these three operations, the degraded urban flow maps can cover various cases of data unreliability. Let d_{max}, d_{min} denote the maximum and minimum values in dataset d and d_{sub} to be $d_{max} - d_{min}$.

- **Data deformation** adopts additive offset, scale transformation, and non-linear conversion, respectively. For additive offset, we introduce a fixed measurement error of $\alpha_1 d_{sub}$ to $r\%$ randomly selected regions. For scale transformation, we randomly scale $r\%$ of the data, denoted as d_{select}, to either $\alpha_2 d_{select}$ or $\alpha_3 d_{select}$ with equal probabilities ($\alpha_2 > 1$ and $\alpha_3 < 1$). For non-linear conversion, we use a non-linear function to transform d_{select} to $t(d_{select})$.
- **Data missing** consists of random region missing and time slot missing. For random region missing, we randomly set the flow value of $\beta_1\%$ regions to 0. As for time slot missing, we randomly select K time intervals in half a day duration (e.g., traffic flow data from 8:00 to 20:00) across the entire dataset. Within these selected intervals, we set the flow values in the $\beta_2\%$ region to 0.
- **Adding noise** uses additive Gaussian noise, pretzel noise and Poisson noise with probabilities of {0.5, 0.3, 0.2} to the original data. For the additive Gaussian distribution, we add the noise with mean as 0 and variance as σ_1. For the pretzel noise, we randomly set the data of σ_2 with equal probability as d_{max} or d_{min}. For Poisson noise, we assume that the incidence is σ_3.

In this experiment, we implement two data degradation settings, degraded-A and degraded-B with the parameters setting as follows. The parameters r, α_1, α_2, and α_3 for data deformation are set to {10, 0.05, 1.05, 0.95} and {20, 0.10, 1.10, 0.90} for the two settings, respectively. The parameters β_1, β_2 and K of the data missing operation are set to {15, 15, 10} and 30, 30, 20}, respectively. The parameters σ_1, σ_2, σ_3 of the adding noise operation are set to {$0.1d_{sub}$, 5%, $0.1d_{sub}$} and {$0.1d_{sub}$, 10%, $0.1d_{sub}$}, respectively.

5.1.3 Baselines. We compare our model against the following six baselines, including statistical based methods, image super-resolution methods, and fine-grained urban flow inference methods.

- **Mean partition (Mean)** distributes the flow volume of a coarse-grained flow region evenly to $n \times n$ fine-grained regions, where n is the upscaling factor. For instance, if a coarse-grained region has crowd flow value as 4 and we need to partition it to 4 fine-grained small regions, each fine-grained region will have the same flow volume as 1.
- **Historical Average (HA)** calculates the average of historical urban flows as the prediction, and then evenly distributes the coarse-grained urban flows into the corresponding 4 fine-grained regions.
- **VDSR[5]** utilizes a deep convolutional network with a depth of 20 to effectively learn the contextual features on large image regions by cascading small filters multiple times to conduct image super-resolution.
- **SRResNet[6]** is a generative adversarial network (GAN) used for image super-resolution (SR). It introduces a perceptual loss function that comprises an adversarial loss and a content loss, enabling better restoration of high-frequency details in the data.
- **UrbanPy[8]** is a recent state-of-the-art fine-grained urban flow inference modele which employs a pyramid architecture containing multiple components. Each component functions as an atomic upsampler for a small scale, which contains an external factor fusion net, an inference network, a proposal net and a correction net.
- **MT-CSR[10]** conducts urban traffic completion and super-resolution simultaneously under a multi-task learning framework. It employs joint training and end-to-end optimization strategies, and considers the local geographical and global semantic correlations within the data.

5.1.4 Evaluation Metrics. We evaluate the inference performance in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), which are defined as follows,

评价指标

$$MAE = \frac{1}{T}\sum_{t=0}^{T}|\mathcal{X}^t - \mathcal{Y}^t|, \tag{9}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=0}^{T}(\|\mathcal{X}^t - \mathcal{Y}^t\|_F^2)}, \tag{10}$$

$$MAPE = \frac{100\%}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left|\frac{\mathcal{X}^t - \mathcal{Y}^t}{\mathcal{Y}^t}\right|, \tag{11}$$

where $\mathcal{X}^t$ is the inferred urban flow spatial map at time t and $\mathcal{Y}^t$ is the corresponding ground truth.

5.2 Comparison Result and Analysis

Table 2 shows the performance comparison among different methods over the two datasets. To more extensively evaluate our model, we conduct the comparison over the original dataset, the dataset with 25% data missing, the dataset with 65% data missing and the dataset with two data degradation settings (degradation-A and degradation-B). The best results are highlighted with boldfont, and the second best results are underlined for a clear comparison.

The results show that our model achieves optimal performance on almost all datasets in most cases. Most of the models have good performance on the original dataset. However, the gap between

Table 2: The performance comparison of different methods over BJTaxi and NYCTaxi under the original data, 25% missing data, 65% missing data, degraded-A and degraded-B settings.

Model			Mean	HA	VDSR	SRResNet	UrbanPy	MT-CSR	DiffUFlow	Improve
BJTaxi	original data	MAE	10.79	1.64	1.65	1.74	1.47	1.62	**1.42**	2.00%
		RMSE	17.87	2.69	2.71	2.78	2.37	2.59	**2.39**	-0.97%
		MAPE	1.12	0.23	0.24	0.26	0.17	0.19	**0.17**	0.00%
	25% data missing	MAE	12.50	2.20	2.13	2.13	1.76	1.59	**1.48**	6.92%
		RMSE	20.51	2.94	2.98	3.09	2.79	2.67	**2.42**	9.36%
		MAPE	1.29	0.33	0.32	0.35	0.26	0.21	**0.17**	17.92%
	65% data missing	MAE	16.13	3.91	2.97	3.01	2.09	1.89	**1.62**	14.39%
		RMSE	25.80	5.94	4.25	4.41	3.32	2.93	**2.59**	11.73%
		MAPE	1.30	0.58	0.42	0.49	0.33	0.28	**0.23**	18.98%
	degraded-A	MAE	18.92	18.71	3.88	3.97	2.98	2.93	**2.35**	19.80%
		RMSE	24.09	21.46	5.79	6.12	4.76	4.78	**3.78**	20.59%
		MAPE	1.87	1.72	0.56	0.62	0.45	0.43	**0.27**	37.21%
	degraded-B	MAE	19.12	18.44	4.09	4.22	3.72	3.88	**2.48**	33.33%
		RMSE	25.37	24.15	6.23	6.48	5.71	5.74	**3.87**	32.22%
		MAPE	1.95	1.91	0.65	0.71	0.54	0.56	**0.32**	40.74%
NYCTaxi	original data	MAE	0.97	0.64	0.58	0.61	0.51	0.55	**0.50**	1.96%
		RMSE	1.85	1.09	0.96	0.97	0.87	0.88	**0.84**	3.45%
		MAPE	0.69	0.47	0.44	0.46	0.37	0.41	**0.35**	4.43%
	25% data missing	MAE	1.10	1.04	0.87	1.02	0.61	0.57	**0.52**	8.77%
		RMSE	2.16	1.98	1.78	1.99	0.99	0.92	**0.87**	5.43%
		MAPE	0.68	0.59	0.56	0.61	0.41	0.39	**0.38**	3.48%
	65% data missing	MAE	1.36	1.30	1.47	1.48	0.69	0.62	**0.58**	6.45%
		RMSE	2.73	2.64	2.83	2.97	1.15	1.07	**0.94**	12.15%
		MAPE	0.67	0.62	0.67	0.72	0.46	0.44	**0.40**	9.55%
	degraded-A	MAE	1.87	1.82	0.99	1.11	0.84	0.92	**0.64**	23.81%
		RMSE	6.02	6.17	1.68	1.84	1.41	1.53	**1.14**	19.15%
		MAPE	1.55	1.64	0.82	0.87	0.70	0.75	**0.43**	38.31%
	degraded-B	MAE	2.14	2.02	1.14	1.30	0.87	0.92	**0.71**	18.39%
		RMSE	6.56	6.45	1.82	2.02	1.56	1.64	**1.34**	14.10%
		MAPE	1.92	1.85	0.91	0.96	0.77	0.82	**0.50**	35.21%

different models is mainly reflected in the missing or degraded data. Conventional statistical methods yield poor results as they simply average a significant amount of erroneous data, resulting in a substantial deviation from the ground truth. VDSR and SRResNet can effectively capture the patterns of traffic flow data, thereby exhibiting a certain resilience to missing and degraded data. Moreover, the fine-grained urban flow inference method leverages traffic-related auxiliary information, resulting in superior performance compared to other baseline models in the case of missing and degraded data.

DiffUFlow shows stronger adaptability on missing or degraded data. Taking the inference performance on BJTaxi dataset as an example, on the severely anomalous data with 65% data missing and degraded-B, DiffUFlow achieves 1.62 and 2.48 MAE results, which are lower by 14.39% and 33.33% compared with the best baseline models, respectively. Furthermore, as the missing data increases from 25% to 65% and the degradation transitions from A to B, the improvement of DiffUFlow increases from 6.92% to 14.39% and from 19.80% to 33.33%, respectively. This indicates that as the task difficulty intensifies, the superiority of DiffUFlow over other models becomes more pronounced. This is because DiffUFlow combines the advantages of several types of models in the baseline. First, DiffUFlow uses self-attentive module instead of a simple convolutional architecture, which makes the model have a more powerful learning capability for complex and long-term temporal correlations. Second, DiffUFlow combines various traffic-related external information, as demonstrated by the results of the ablation experiments in Section 5.3, which contributes to the performance improvement of the model. Finally, the design of the diffusion model is naturally suitable for such data with errors because the process of recovering anomalous data can be considered as part of its reverse denoising process. This makes DiffUFlow performs better in urban flow inference, exhibiting superior performance, robustness, and generalization capabilities in real-life scenarios.

5.3 Ablation Study

To further demonstrate the effectiveness of the proposed components in DiffUFlow, we also compare the full version of DiffUFlow (denoted as DUF) with the following variants.

- **DUF-V1** removes both spatial-temporal hybrid attention network STFormer and semantic features fetcher ELFetcher. By comparing with it, we can test whether the rich unique features of urban flows are useful for fine-grained inference.
- **DUF-V2** drops the spatial-temporal hybrid attention network STFormer. Through comparing with this model, we test whether STFormer can enhance the spatial-temporal

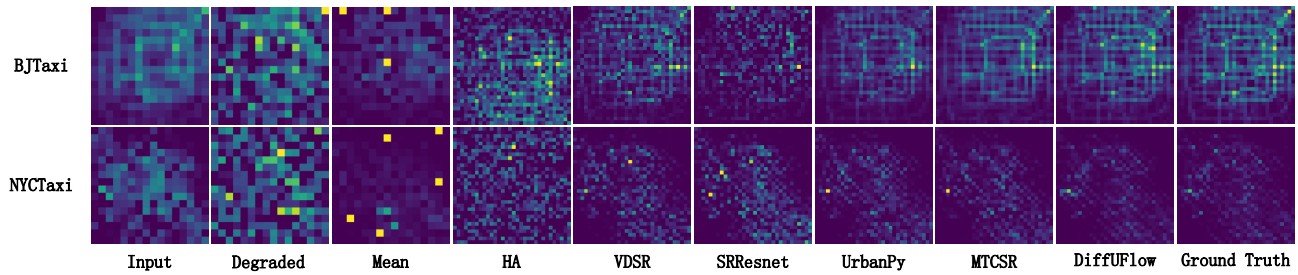

Figure 4: Visualization of the urban flows inference results on BJTaxi at 2015/5/1 12:00 and NYCTaxi at 2015/10/10 12:00. The input size is 16×16 and output size is 32×32.

Table 3: The ablation study result.

		DUF-V1	DUF-V2	DUF-V3	DUF
STFormer		✗	✔	✗	✔
ELFetcher		✗	✗	✔	✔
original	MAE	1.60	1.48	1.54	1.42
	RMSE	2.71	2.51	2.63	2.39
	MAPE	18.34%	17.41%	17.64%	17.27%
25% data missing	MAE	1.74	1.61	1.65	1.48
	RMSE	2.97	2.69	2.84	2.42
	MAPE	21.05%	18.98%	19.54%	17.43%
degraded-A	MAE	3.07	2.49	2.68	2.35
	RMSE	5.23	4.04	4.52	3.78
	MAPE	34.56%	29.74%	31.74%	27.69%

features of historical urban flow maps, and thus improve the model performance.

- **DUF-V3** removes the semantic features fetcher ELFetcher. Through comparing with it, we test whether the semantic features including external factors and land features are useful for our model.

We conduct the ablation study on the BJTaxi dataset under three different scenarios, original data, 25% data missing, and degraded-A setting. The result is shown in Table 3. The result demonstrates that discarding either the STFormer or the ELFetcher leads to a significant performance decline. One can notice that STFormer contributes more significantly to the inference performance. We ascribe the superiority to the spatial-temporal dependencies extracted by the STFormer, which is more informative and important to the urban flow inference process than the external factors. By integrating STFormer and ELFetcher, DiffUFlow can fully capture the spatial-temporal dependencies and the semantic features to better infer the fine-grained urban flows when the data are noisy and incomplete.

5.4 Visualization and Case Study

For an intuitive performance comparison between DiffUFlow and the baselines, we provide a visualized case on BJTaxi dataset at 2015/5/1 12:00 and NYCTaxi dataset at 2015/10/10 12:00 under the degraded-A setting. As shown in Figure 4, we visualize the inference results of all models. One can see that the traditional statistical methods perform poorly due to the unreliable regions caused by noise. HA that utilizes the temporal information performs slightly better than mean partition. VDSR and SRResNet can learn the patterns of traffic flow and generate a blurred outline, but there is still a significant gap compared to the ground-truth. UrbanPy and MT-CSR effectively leverage the auxiliary traffic-related information, which are more robust to noise and generate more promising result than VDSR and MT-CSR. As shown in the visualization result, the inferred fined-grained urban flow map matches the ground truth best. During the inference procedure, DiffUFlow integrates the spatial-temporal dependencies with the semantic features and eliminates the noisy data by reverse diffusion process, resulting in the best performance.

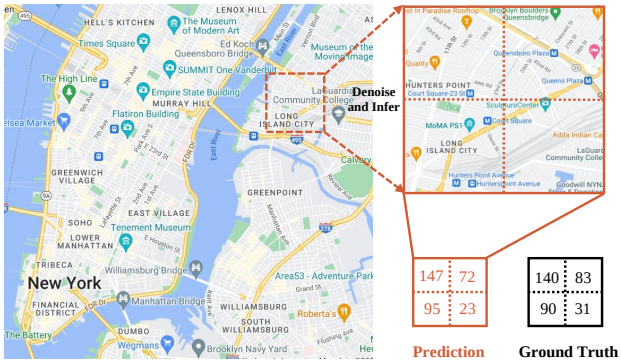

Figure 5: Case study of the inferred urban flows by DiffUFlow in one area of NewYork at 2016/2/20 08:30.am.

In order to further explore the denoising capability of DiffUFlow, we conduct a case study to evaluate the inference performance with noise on predictions. We select the morning of February 20, 2016, at 8:30 am in NewYork and add noise in certain regions. The inference result is shown in Figure 5. It can be observed that the red squares represent the areas where the traffic data was corrupted with noise, increasing the value from 337 to 394. It is worth noting that the predicted sum of urban flow remains at 342, which indicates that the robustness of the model to the noise.

6 CONCLUSION

In this paper, we proposed an diffusion-based model for robust fine-grained urban flow inference with noisy and incomplete data. DiffUFlow addressed two challenges which were significant to robust fine-grained urban flow inference process, including the noisy urban flows, the complex and dynamic spatial-temporal information and conducting robust urban flow super-resolution. Extensive evaluations on two real large datasets showed that the proposed model significantly enhanced the performance and outperformed the state-of-the-art models. Studies and visualizations also confirmed the effectiveness of DiffUFlow.

ACKNOWLEDGEMENT

This research was funded by the National Science Foundation of China (Nos. 62172443 and 62172034), Open Project of Xiangjiang Laboratory (22XJ02002, 22XJ03025), Hunan Provincial Natural Science Foundation of China (No. 2022JJ30053), the Science and Technology Major Project of Changsha (No. kh2202004), the Beijing Natural Science Foundation (No. 4212021), PolyU RIAIoT and RIO (No. BD4A).

REFERENCES

[1] Yuxuan Liang, Kun Ouyang, Lin Jing, Sijie Ruan, Ye Liu, Junbo Zhang, David S Rosenblum, and Yu Zheng. Urbanfm: Inferring fine-grained urban flows. In *Proceedings of KDD*, 2019.
[2] Wenbin Liu, Yongjian Yang, En Wang, and Jie Wu. Fine-grained urban prediction via sparse mobile crowdsensing. In *Proceedings of MASS*, 2020.
[3] Mingxiao Li, Hengcai Zhang, and Jie Chen. Fine-grained dynamic population mapping method based on large-scale sparse mobile phone data. In *Proceeding of MDM*, 2019.
[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2015.
[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of CVPR*, 2016.
[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of CVPR*, 2017.
[7] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
[8] Kun Ouyang, Yuxuan Liang, Ye Liu, Zekun Tong, Sijie Ruan, Yu Zheng, and David S Rosenblum. Fine-grained urban flow inference. *IEEE transactions on knowledge and data engineering*, 2020.
[9] Yuxuan Liang, Kun Ouyang, Yiwei Wang, Ye Liu, Junbo Zhang, Yu Zheng, and David S Rosenblum. Revisiting convolutional neural networks for citywide crowd flow analytics. In *Proceeding of ECML-PKDD*, 2021.
[10] Jiyue Li, Senzhang Wang, Jiaqiang Zhang, Hao Miao, Junbo Zhang, and S Yu Philip. Fine-grained urban flow inference with incomplete data. *IEEE Transactions on Knowledge and Data Engineering*, 2022.
[11] Xueyan Yin, Genze Wu, Jinze Wei, Yanming Shen, Heng Qi, and Baocai Yin. Deep learning on traffic prediction: Methods, analysis, and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 2021.
[12] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of AAAI*, 2017.
[13] Senzhang Wang, Jiannong Cao, Hao Chen, Hao Peng, and Zhiqiu Huang. Seqst-gan: Seq2seq generative adversarial nets for multi-step urban crowd flow prediction. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 6(4):1–24, 2020.
[14] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.
[15] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceeding of AAAI*, 2020.
[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI*, 2015.
[17] Senzhang Wang, Jiannong Cao, and S Yu Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2022.
[18] Bowen Du, Hao Peng, Senzhang Wang, Md Zakirul Alam Bhuiyan, Lihong Wang, Qiran Gong, Lin Liu, and Jing Li. Deep irregular convolutional residual lstm for urban traffic passenger flows prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):972–985, 2019.
[19] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceeding of AAAI*, 2017.
[20] Haoxing Lin, Rufan Bai, Weijia Jia, Xinyu Yang, and Yongjian You. Preserving dynamic attention for long-term spatial-temporal prediction. In *Proceedings of KDD*, 2020.
[21] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
[22] Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 2017.
[23] Weiqi Chen, Ling Chen, Yu Xie, Wei Cao, Yusong Gao, and Xiaojie Feng. Multi-range attentive bicomponent graph convolutional network for traffic forecasting. In *Proceeding of AAAI*, 2020.
[24] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Proceeding of NeurIPS*, 2020.
[25] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proceeding of NeurIPS*, 2021.
[26] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022.
[27] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of NeurIPS*, 2021.
[28] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.
[29] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Proceedings of NeurIPS*, 2021.
[30] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceeding of CVPR*, 2022.
[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceeding of CVPR*, 2022.
[32] Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv preprint arXiv:2208.08664*, 2022.
[33] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.
[34] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv*, 2022.
[35] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceeding of CVPR*, 2022.
[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of NeurIPS*, 2020.
[37] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021.
[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceeding of CVPR*, 2016.