Overcoming Forgetting in Fine-Grained Urban Flow Inference via Adaptive Knowledge Replay

Haoyang Yu,[*] **Xovee Xu,**[*] **Ting Zhong,**[†] **Fan Zhou** [†]

University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
haoyang.yu417@outlook.com, xovee@live.com, {zhongting, fan.zhou}@uestc.edu.cn

Abstract

Fine-grained urban flow inference (FUFI) problem aims at inferring the high-resolution flow maps from the coarse-grained ones, which plays an important role in sustainable and economic urban computing and traffic management. Previous models addressed the FUFI problem from spatial constraint, external factors, and memory cost. However, utilizing the new urban flow maps to calibrate the learned model is very challenging due to the "catastrophic forgetting" problem and is still under-explored. In this paper, we make the first step in FUFI and present CUFAR – Continual Urban Flow inference with Adaptive knowledge Replay – a novel framework for inferring the fine-grained citywide traffic flows. Specifically, (1) we design a spatial-temporal inference network that can extract better flow map features from both local and global levels; (2) then we present an adaptive knowledge replay (AKR) training algorithm to selectively replay the learned knowledge to facilitate the learning process of the model on new knowledge without forgetting. In addition, we also propose a knowledge discriminator to avoid "negative replaying" issue introduced by noisy urban flow maps. Extensive experiments on four large-scale real-world FUFI datasets demonstrate that our proposed model consistently outperforms strong baselines and effectively mitigates the forgetting problem. Source code is available at: https://github.com/PattonYu/CUFAR.

1 Introduction

Fine-grained urban flow analysis, prediction, and inference are important applications of smart city development and urban computing. They have been used for traffic management and urban transportation planning (Cleophas et al. 2019; Liang et al. 2022b; Zheng et al. 2014; Liang et al. 2019), benefiting from the fast urbanization, vast data generated from IoT devices, and the new computing technologies in recent years (Zhong et al. 2022; Yu et al. 2022). However, extensive resources such as electricity and manpower are consumed for deploying and maintaining the system. Fine-grained urban flow inference (FUFI), which tries to infer the high-resolution flow map from the corresponding coarse-grained one, is proposed as an important step toward an environmental-friendly and sustainable urban traffic system.

UrbanFM (Liang et al. 2019) is the first work formulates the FUFI problem and proposes a distributional upsampling module and an external factor fusing subnet for tackling the problem. Subsequent works improve UrbanFM from several important aspects, e.g., the spatial constraint, external factors, and memory cost. Specifically, FODE (Zhou et al. 2020) and UrbanODE (Zhou et al. 2021) – based on neural ordinary differential equations (ODEs) – are proposed to address the numerical instability problem in FUFI by an affine coupling layer and a pyramid attention network. MT-CSR (Li et al. 2022) addresses the FUFI problem with incomplete urban flow map. DeepLGR (Liang et al. 2020) revisits the limitations of convolutional neural network (CNN) and tries to learn global spatial dependencies and local feature representations of the flow dynamics. UrbanPy (Ouyang et al. 2022) is the state-of-the-art FUFI model which extends UrbanFM by proposing a cascading strategy based on the pyramid architecture, a propose-and-correct component, and a new distribution loss.

Motivations. Despite the promising results achieved in prior works, several potential improvements are worth exploring. First, although existing works have designed global-local architectures such as pyramid mechanism (Ouyang et al. 2022; Zhou et al. 2021), global-local context module (Liang et al. 2020), and Transformer (Zhou, Zhou, and Liu 2021; Liang et al. 2022a) to learn the long-range dependencies between local regions at different granularity, they are still inefficient in modeling the spatial relations of urban flows while also considering the temporal flow dynamics. Second, prior methods learn each FUFI dataset in isolation and retrain the entire model with the newly obtained fine-grained urban flow maps, leaving the previous data unexploited. One straightforward solution is to train all the data at once. However, it has two drawbacks: (1) noise data may be introduced from the older flow maps that have different flow distributions; (2) the computation overhead becomes unaffordable as time goes on. Another solution is to continually fine-tune the trained model from previous data on the new data, which is efficient and feasible to take advantage of the learned urban flow knowledge. However, fine-tuning directly on the new data is very prone to the "catastrophic forgetting" problem – i.e., much of the learned knowledge is overridden upon learning the new knowledge

[*] These authors contributed equally.

[†] Corresponding author.

– mainly due to the parameter-updating mechanism (e.g., back-propagation), resulting in the model less generalized and less robust. This is also a result of "stability-plasticity" dilemma (Carpenter and Grossberg 1987).

Present Work. We present CUFAR: Continual Urban Flow inference with Adaptive knowledge Replay, as a novel way of inferring the fine-grained flow map with the help of previously learned knowledge. Specifically, we design a simple yet effective inference network that extracts spatial-temporal flow map features from both local and global perspectives, enabling the model to infer more accurate flow distributions. Then we propose a general adaptive knowledge replay (AKR) training algorithm to continually and selectively replay the old knowledge to facilitate the learning process of the model on new data while also overcoming the "catastrophic forgetting" problem. Moreover, we design an adaptive knowledge discriminator to measure the flow distribution difference before and after the knowledge replaying, which helps the model mitigate the "negative replaying" issue that may occur if noisy data are introduced.

Extensive experiments (including ablation study and visualization) on four large-scale real-world urban flow datasets demonstrate the effectiveness and robustness of CUFAR over strong FUFI baselines. We have the following notable findings: (1) the proposed spatial-temporal inference network uniformly improved the FUFI performance on four datasets, which has a better capability for learning expressive flow map features. (2) the designed AKR training algorithm successfully alleviated the "catastrophic forgetting" problem in continual FUFI and consequently, improved the inference performance. It is worth noting that all baselines equipped with AKR have better performance. (3) interestingly, on TaxiBJ-P4 dataset, we both observed the "negative replaying" and "overfitting" if the proposed knowledge discriminator and AKR are removed, respectively. (4) compared to the *joint* protocol, i.e., training all data at once, our approach is efficient and even outperforms *joint*. We speculate this deficiency of *joint* is due to the noisy samples introduced from previous data. The above findings verify our motivation and show that utilizing the prior knowledge (properly) to facilitate the learning process of current knowledge is a promising way toward a robust and sustainable urban transportation system.

2 Problem Formulation

We aim to infer the citywide fine-grained flow map from the coarse-grained one. Given a city of interest, we divide the city's map M into grid-cells. For each cell, we record its traffic flows $x_{ij} \in \mathbb{R}_+$ every τ minutes. The overall traffic flows of M are denoted as $\mathbf{X}$. Following exiting works (Liang et al. 2019; Ouyang et al. 2022), the FUFI problem and its spatial constraint are defined as:

Definition 1. Fine-Grained Urban Flow Inference: Given a coarse-grained map $\mathbf{X}_{cg} \in \mathbb{R}_+^{H\times W}$, the FUFI problem is to infer the corresponding fine-grained flow map $\mathbf{X}_{fg} \in \mathbb{R}_+^{NH\times NW}$, here N is an upscaling factor.

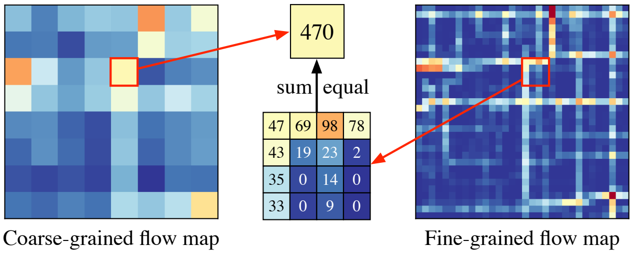

Figure 1: Spatial constraint between coarse- and fine-grained flow maps in a local area of Beijing city.

Definition 2. Spatial Constraint: Different from image super-resolution, FUFI problem has to obey a spatial constraint that the cell flow x_{ij} of the coarse-grained map is strictly equal to the sum of flows in the corresponding $N \times N$ cells of the fine-grained map, i.e.:

$$x_{ij,cg} = \sum_{i'j'} x_{i'j',fg} \quad s.t. \left\lfloor \frac{i'}{N} \right\rfloor = i, \left\lfloor \frac{j'}{N} \right\rfloor = j, \qquad (1)$$

where $i = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, W$. An illustration of the spatial constraint is depicted in Figure 1.

In traditional FUFI problem settings (Liang et al. 2019; Li et al. 2020), the learning model is retrained entirely every time new urban data comes in, which belongs to the offline learning paradigm. Online learning, on the contrary, dynamically and sequentially learns new data patterns, enabling an efficient and sustainable prediction model that is also the target of the FUFI problem. However, online learning algorithms such as continual learning and fine-tuning may face the challenge of catastrophic forgetting when adapting to new data/tasks[1]. In this work, we study the continual FUFI problem, i.e., given a sequence of urban flow datasets ordered over time, we learn new knowledge with the help of old knowledge, but without forgetting the old knowledge.

3 Methodology

We now illustrate the CUFAR methodology which contains two critical components: (1) an inference network consisted of two feature extractors for learning the spatial-temporal relations of urban flow maps from both global and local levels; (2) a specifically designed continual algorithm that combines the experience replay strategy with an adaptive knowledge discriminator. The framework of the inference network is depicted in Figure 2 and the continual algorithm is described in Algorithm 1.

3.1 Spatial-Temporal Inference Network

Spatial Relation Extraction. Urban road topological structures and the corresponding temporal flow dynamics are extremely complex and cannot be parsed with simple rules. With the help of CNNs, previous FUFI methods (Liang et al. 2019; Zhou et al. 2020, 2021) learn global

[1] When there is no ambiguity, we interchangeably use data and task throughout the paper.

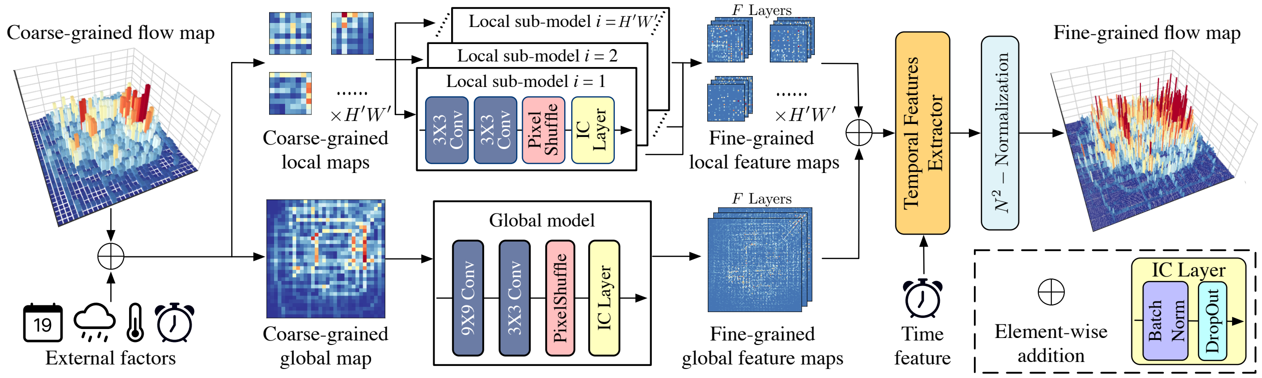

Figure 2: The framework of the proposed inference network. It first extracts global-local map features from the coarse-grained flow map along with external factors such as weather and date. Then extracted flow map features are combined with temporal features and a N^2-Normalization layer to infer the final fine-grained flow map.

feature maps with shared kernel weights to infer the fine-grained flow map. As a result, these methods are inefficient to model the local area flow dynamics. To address this hurdle, we design a spatial relation extraction module to partition the flow map into smaller local regions and use several standalone sub-models to separately infer the upscaled local maps. Specifically, for $H'W'$ sub regions, each sub-model is consisted of two convolution layers followed by a PixelShuffle layer and an independent component (IC) layer. The first convolution layer has F filters (3x3) and the second has FN^2 filters, recall that N is the upscale factor. The PixelShuffle layer upscales the coarse-grained feature maps $\mathbf{H}_{cg}^{\text{local},i} \in \mathbb{R}^{\frac{H}{H'} \times \frac{W}{W'} \times FN^2}$ to fine-grained ones $\mathbf{H}_{fg}^{\text{local},i} \in \mathbb{R}^{\frac{NH}{H'} \times \frac{NW}{W'} \times F}$. The IC layer (Chen et al. 2019) is composed of a batch normalization layer and a dropout layer. It can whiten the mutual information and correlation coefficients between network neurons and accelerate the speed of convergence. The global model is similar to the local sub-model but has a larger filter size (9x9) in the first convolution layer. Its output is the global fine-grained feature maps $\mathbf{H}_{fg}^{\text{global}} \in \mathbb{R}^{NH \times NW \times F}$. We then restore the undivided feature maps $\mathbf{H}_{fg}^{\text{local}} \in \mathbb{R}^{NH \times NW \times F}$ from local feature maps. At last, we concatenate the obtained feature maps as the input of the temporal feature extractor: $\mathbf{H}_{fg} = \left[\mathbf{H}_{fg}^{\text{global}}; \mathbf{H}_{fg}^{\text{local}}\right] \in \mathbb{R}^{NH \times NW \times 2F}$.

Temporal Features Extraction. Various external conditions such as *weather* and *date* have an influence on the distribution of citywide urban flows. Among them, time span is closely related to flow volumes but largely ignored or dealt as a normal external condition. To remedy this, we propose a temporal feature extraction module based on convolutional sequences to capture the influence of time factor on the traffic flow distribution. Specifically, we build K independent convolutional layers (F filters, 3x3), each account for a specific time span in a day. The weights of these layers are non-shared. Then the fine-grained feature maps $\mathbf{H}_{fg}$ are fed into the corresponding convolutional layer followed by a N^2-Normalization layer to obtain the desired fine-grained flow map $\widetilde{\mathbf{X}}_{fg} \in \mathbb{R}^{NH \times NW}$.

External Factors and N^2-Normalization. As same as the one in UrbanFM (Liang et al. 2019), we adopt several embedding layers to transform the external factors (except the time span factor) into low-dimensional vectors and then use dense layers to reshape the vectors into a coarse-grained feature map. Thereby, the input of our model is the concatenation of the coarse-grained flow map, external feature map, and time span feature map. To obey the spatial constraint required by FUFI problem, existing methods often adopt the N^2-Normalization as model's last layer instead of adding new losses. In CUFAR, we also use this normalization trick. The details of the external factors and N^2-Normalization can be found in UrbanFM (Liang et al. 2019).

Optimization. The training objective of CUFAR is the widely used mean squared error (MSE) between the inferred flow map and the ground truth:

$$\mathcal{L} = ||\widetilde{\mathbf{X}}_{fg} - \mathbf{X}_{fg}||^2. \tag{2}$$

Next, we illustrate how we adaptively and continually learn new knowledge without forgetting the old knowledge.

3.2 Adaptive Knowledge Replay

When we use the newly obtained urban flow maps to calibrate the trained model in FUFI (we denote this process as learning on a task), e.g., fine-tuning on the new data, the learned knowledge tend to be overridden by the new knowledge due to the parameter-updating mechanism (e.g., backpropagation), resulting catastrophic forgetting that lowers the inference performance. In this work, we propose an adaptive knowledge replay (AKR) training method that continually and selectively replays the old knowledge to help the learning process of the new task. AKR consists of a memory buffer and an adaptive knowledge discriminator.

Selective Memory Buffer. Prior works (Lopez-Paz and Ranzato 2017; Buzzega et al. 2020; Riemer et al. 2019) show that one effective way to overcome catastrophic forgetting is the experience replay, which saves the old data in a memory buffer and does not constrain the optimization process.

Inspired by experience replay, we use a memory buffer $\mathcal{M}$ to reserve the learned urban flow maps from previous tasks, which has a max size S. In every training iteration, we randomly sample a buffer mini-batch $B_{\mathcal{M}}$ from $\mathcal{M}$. The sampled $B_{\mathcal{M}}$ is then merged with the original training batch B. Different from existing experience replay-based methods (Buzzega et al. 2020; Riemer et al. 2019), our designed selective memory buffer (1) does not retrospect current training data and (2) contains a sub-memory buffer $\mathcal{M}_{\text{sub}}$ of size $S/2$ for storing recent data on the new task. Specifically, during the training on the first task, buffer $\mathcal{M}$ is filled and then updated by the reservoir sampling algorithm (Vitter 1985). During the training of subsequent tasks, the sub-memory buffer $\mathcal{M}_{\text{sub}}$ replaces the role of the $\mathcal{M}$, i.e., we update $\mathcal{M}_{\text{sub}}$ with the new data and replay the old data from $\mathcal{M}$. Every time we train on new task, half of the data in $\mathcal{M}$ is randomly replaced by the data (from the last task) in $\mathcal{M}_{\text{sub}}$. Therefore, the buffer $\mathcal{M}$ has more recent samples.

Adaptive Knowledge Discriminator. When replaying past data from the memory buffer $\mathcal{M}$, if the distribution of replayed urban flow maps greatly differs from the original distribution, "negative replaying" may occur. To mitigate this deficiency, we further introduce a maximum mean discrepancy (MMD) (Gretton et al. 2012) to measure the similarity between the replayed distribution and original distribution. MMD is often used in domain adaptation and transfer learning to measure the domain distance and constrain the representation space (Liu et al. 2020; Bińkowski et al. 2018; Arbel et al. 2018). We use MMD as a distance discriminator to prevent our model being affected by the out-of-distribution flow maps (e.g., unusual traffic flow distribution due to road accidents or lockdowns). Given two distributions – in our case the replayed samples $\mathbf{X}_{cg}^{\text{replay}}$ from the merged batch and samples $\mathbf{X}_{cg}^{\text{ori}}$ in the original batch – the MMD distance is defined as:

$$d_{\text{MMD}}(\mathbf{X}_{cg}^{\text{replay}}, \mathbf{X}_{cg}^{\text{ori}}) = \sup_{f \in \mathcal{H}} \left(\mathrm{E}[f(P)] - \mathrm{E}[f(Q)] \right), \tag{3}$$

where $P \sim \mathbf{X}_{cg}^{\text{replay}}$, $Q \sim \mathbf{X}_{cg}^{\text{ori}}$, and $\mathcal{H}$ indicates reproducing kernel Hilbert space (RKHS). We define two kernel embeddings $\mu_p := E[\mathcal{K}(\cdot, P)]$ and $\mu_q := E[\mathcal{K}(\cdot, Q)]$, here $\mathcal{K}(\cdot,\cdot) \in \mathcal{H}$. Then the MMD distance can be derived as:

$$d_{\text{MMD}}^2(\mathbf{X}_{cg}^{\text{replay}}, \mathbf{X}_{cg}^{\text{ori}}) = \left[\sup_{\|f\|_{\mathcal{H}} \leqslant 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right]^2 \tag{4}$$

$$\leqslant \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \, \|f\|_{\mathcal{H}}^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \, .$$

Since the above equation cannot be computed directly, we expand the kernel function and draw i.i.d. samples $P = \{p_i\}_{i=1}^m$ from $\mathbf{X}_{cg}^{\text{replay}}$ and $Q = \{q_j\}_{j=1}^n$ from $\mathbf{X}_{cg}^{\text{ori}}$. Then the squared MMD can be estimated as follows (Gretton et al. 2012; Bińkowski et al. 2018):

$$d_{\text{MMD}}^2(P, Q) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k\left(p_i, p_j\right) \tag{5}$$

$$+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k\left(q_i, q_j\right) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k\left(p_i, q_j\right).$$

Algorithm 1: Adaptive Knowledge Replay (AKR)

```
Input: A sequence of tasks $\{T_1, T_2, \dots\}$ containing coarse- and
    fine-grained flow maps $\mathbf{X}_{cg}$ and $\mathbf{X}_{fg}$, buffer $\mathcal{M}$ and $\mathcal{M}_{\text{sub}}$.
$\mathcal{M} \leftarrow \varnothing, \mathcal{M}_{\text{sub}} \leftarrow \varnothing$;
for $T_i$ in $\{T_1, T_2, \dots\}$ do
    if $T_1$ then
        for sampled mini-batch $B = \{(\mathbf{X}_{cg}^j, \mathbf{X}_{fg}^j)\}_{j=1}^{|B|}$ do
            Fill $\mathcal{M}$ using the reservoir sampling algorithm;
            for $\mathbf{X}_{cg}^j \in B$ do
                $\widetilde{\mathbf{X}}_{fg}^j \leftarrow \text{Model}(\mathbf{X}_{cg}^j)$;
            end for
            $\theta \leftarrow \text{MSELoss}(\{\widetilde{\mathbf{X}}_{fg}^j, \mathbf{X}_{fg}^j\}_{j=1}^{|B|})$;        ▷ Updating
        end for
    else                                        ▷ Starting to replay knowledge
        if $\mathcal{M}_{\text{sub}} \neq \varnothing$ then
            Replace half of $\mathcal{M}$ with $\mathcal{M}_{\text{sub}}$ and set $\mathcal{M}_{\text{sub}} = \varnothing$;
        end if
        for sampled mini-batch $B = \{(\mathbf{X}_{cg}^{\text{ori},j}, \mathbf{X}_{fg}^{\text{ori},j})\}_{j=1}^{|B|}$ do
            Fill $\mathcal{M}_{\text{sub}}$ using the reservoir sampling algorithm;
            Sample $B_{\mathcal{M}}$ from the buffer $\mathcal{M}$;
            $B^{\text{replay}} = \{(\mathbf{X}_{cg}^{\text{replay},j}, \mathbf{X}_{fg}^{\text{replay},j})\}_j \leftarrow B \cup B_{\mathcal{M}}$;
            $\alpha \leftarrow d_{\text{MMD}}^2(B^{\text{replay}}, B), \theta_0 \leftarrow \theta$;
            for $\mathbf{X}_{cg}^{\text{replay},j} \in B^{\text{replay}}$ do
                $\widetilde{\mathbf{X}}_{fg}^{\text{replay},j} = \text{Model}(\mathbf{X}_{cg}^{\text{replay},j})$;
            end for
            $\theta_1 \leftarrow \text{MSELoss}(\{\widetilde{\mathbf{X}}_{fg}^{\text{replay},j}, \mathbf{X}_{fg}^{\text{replay},j}\}_{j=1}^{B^{\text{replay}}})$;
            $\theta \leftarrow \theta_0 + \alpha * (\theta_1 - \theta_0)$;                ▷ Updating
        end for
    end if
end for
```

The range of d_{MMD}^2 is $[0, +\infty]$. We then normalize the distance into $(0, 1]$ by:

$$\alpha = 2 - \frac{2}{1 + e^{-d_{\text{MMD}}^2}}. \tag{6}$$

The larger the α, the higher the similarity, and $\alpha = 1$ indicates that the two sample groups are identical. When replaying the learned knowledge on new tasks, for each training iteration, let θ_0 be the initial weights, θ_1 be the trained weights, the final model weights θ are updated as:

$$\theta = \theta_0 + \alpha * (\theta_1 - \theta_0). \tag{7}$$

The overall training process of the proposed adaptive knowledge replay is sketched in Algorithm 1.

4 Experiments

We now evaluate the effectiveness of our proposed spatial-temporal inference network and adaptive knowledge replay on continual FUFI problem. Experiments are conducted on four real-world taxi traffic datasets collected continuously for four years (2013 to 2016) in Beijing (Liang et al. 2019). We denote the four datasets as TaxiBJ Task-1 to Task-4.

Baselines. We compare CUFAR with the following ten methods on FUFI problem. They include five single image super-resolution approaches: SRCNN (Dong et al. 2015), VDSR (Kim, Lee, and Lee 2016), ESPCN (Shi et al. 2016), SRResNet (Ledig et al. 2017) and DeepSD (Vandal et al.

Method	**Task-1**			**Task-2**			**Task-3**			**Task-4**		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
SRCNN	18.464	2.491	0.714	21.270	2.681	0.689	23.184	2.829	0.727	14.730	2.289	0.665
ESPCN	17.690	2.497	0.732	20.875	2.727	0.732	22.505	2.862	0.773	13.898	2.228	0.711
VDSR	17.297	2.213	0.467	21.031	2.498	0.486	22.372	2.548	0.461	13.351	1.978	0.411
DeepSD	17.272	2.368	0.614	20.738	2.612	0.621	22.014	2.739	0.682	15.031	2.297	0.652
SRResNet	17.338	2.457	0.713	20.466	2.660	0.688	21.996	2.775	0.717	13.446	2.189	0.637
UrbanFM	16.372	2.066	0.335	19.548	2.284	0.328	21.243	2.398	0.336	12.744	1.850	0.311
DeepLGR	17.125	2.103	0.339	21.217	2.386	0.350	23.563	2.497	0.351	13.390	1.916	0.345
FODE	16.473	2.142	0.403	19.884	2.377	0.395	21.425	2.490	0.417	12.840	1.947	0.396
UrbanODE	16.342	2.135	0.406	19.648	2.357	0.394	21.177	2.460	0.408	12.668	1.929	0.391
UrbanPy	16.082	2.026	0.329	19.025	2.232	0.318	20.810	2.333	0.313	12.336	1.810	0.304
CUFAR	**14.991**	**1.952**	**0.306**	**18.259**	**2.186**	**0.301**	**19.309**	**2.243**	**0.289**	**11.681**	**1.758**	**0.288**

Table 1: Inference performance on four TaxiBJ datasets when using the *single-task* protocol. Results of the five single image super-resolution baselines are from (Liang et al. 2019). The best results are marked in **bold**.

	Method	**Task-2**			**Task-3**			**Task-4**		
		MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
fine-tune	UrbanFM	19.162	2.257	0.322	20.499	2.341	0.325	12.285	1.814	0.314
	DeepLGR	20.571	2.336	0.334	21.845	2.427	0.345	12.820	1.858	0.318
	FODE	19.251	2.323	0.379	20.511	2.410	0.387	12.414	1.895	0.369
	UrbanODE	19.070	2.302	0.371	20.275	2.375	0.372	12.182	1.862	0.361
	UrbanPy	18.822	2.208	0.317	20.117	2.293	0.314	12.088	1.800	0.307
	CUFAR	**17.746**	**2.151**	**0.293**	**18.915**	**2.219**	**0.287**	**11.486**	**1.745**	**0.290**
continual	UrbanFM	18.477	2.215	0.312	19.809	2.290	0.314	11.919	1.778	0.302
	DeepLGR	19.202	2.292	0.342	19.892	2.331	0.330	11.977	1.819	0.314
	FODE	18.799	2.297	0.370	20.012	2.369	0.373	11.997	1.852	0.359
	UrbanODE	18.735	2.289	0.374	19.779	2.340	0.361	11.924	1.836	0.352
	UrbanPy	18.286	2.193	0.311	19.503	2.264	0.314	11.958	1.787	0.304
	CUFAR	**17.616**	**2.141**	**0.292**	**18.840**	**2.213**	**0.285**	**11.420**	**1.735**	**0.283**

Table 2: Performance comparison between *fine-tune* and *continual* protocols. All models are initially trained on Task-1. Then each model is fine-tuned on new tasks. *Continual* indicates our designed AKR is applied. Best results are in **bold**.

2017); and five state-of-the-art FUFI approaches: UrbanFM (Liang et al. 2019), DeepLGR (Liang et al. 2020), FODE (Zhou et al. 2020), UrbanODE (Zhou et al. 2021) and UrbanPy (Ouyang et al. 2022).

Training Protocols. We use four training protocols. *Single-task* learns each task in isolation. *Joint* learns previous tasks and new task at once, which costs a lot of computations when there are many tasks. *Fine-tune* and *continual* learn new task with the help of previous task(s), but do not require retraining on them.

Implementation. We use commonly used FUFI metrics including mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). All experiments are conducted on RTX 3090 with PyTorch. The optimizer is Adam, learning rate is $1e^{-4}$, filter size F is 128, temporal conv layers K is 15 (hourly from 9AM to 12PM), memory buffer size S is 1,000, batch B and $B_{\mathcal{M}}$'s sizes are 16 and 2, respectively. The resolution of the flow map $\mathbf{X}_{fg}$ is 128x128, the upscaling factor N is 4.

4.1 Evaluation Results

Spatial-Temporal Inference Network. Table 1 shows the inference performance of our model and ten baselines on four tasks using the *single-task* protocol, i.e., we infer the fine-grained urban flow without fine-tuning or knowledge-replaying on the new task. We can observe CUFAR achieves the best results through all metrics on all tasks, which verifies the effectiveness of our designed spatial-temporal inference network.

Adaptive Knowledge Replay. Next, we evaluate the proposed AKR training algorithm and see if we can mitigate catastrophic forgetting while also improving the inference performance. Since image super-resolution methods performed poorly, we omit them in the remaining experiments. We apply AKR on all FUFI methods (denoted as *continual*) and the results are shown in Table 2. The comparison shows that *continual*-based models consistently outperform *fine-tune*-based models, supporting our motivation that overcoming catastrophic forgetting is vital for the FUFI problem. In addition, *fine-tune* and *continual* both considerably

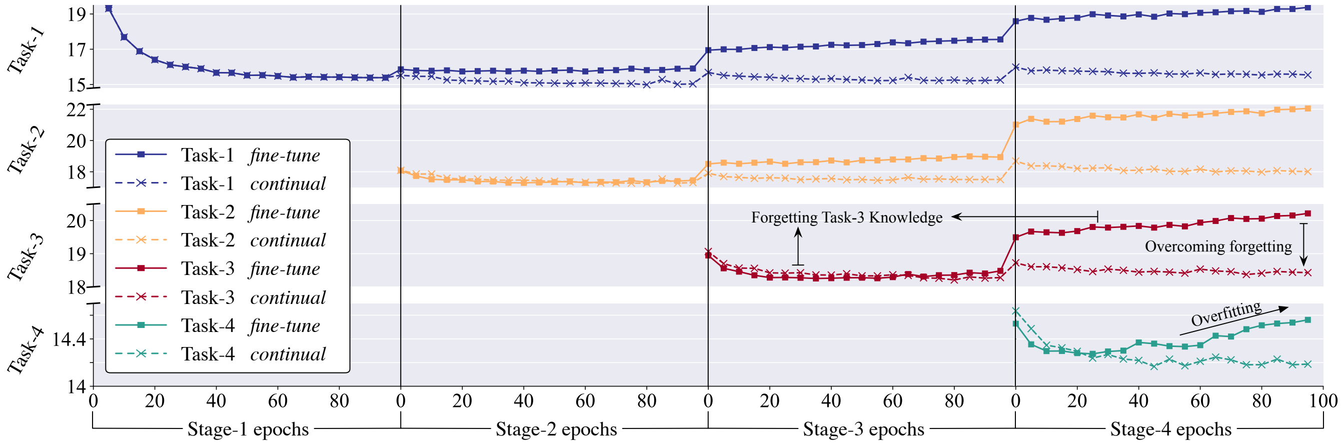

Figure 3: Visualization of catastrophic forgetting phenomenon. Lines are validation losses (MSE). Each row represents a task. Each column represents a training stage. Each stage we train the model on current task (by *fine-tune* or *continual* protocol) and validate the model on current task and (if any) previous tasks. As we continuously learn on new tasks, the performance of the *fine-tune* model on previous tasks drops severely due to catastrophic forgetting, while *continual* largely alleviates this issue.

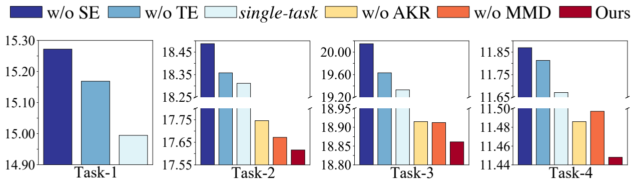

Figure 4: Ablation study results in terms of MSE.

outperform *single-task* (compare the results of Table 1 and Table 2), which suggests that utilizing the knowledge from previous tasks is an effective way to boost the performances of models on new tasks. Interestingly, UrbanFM with *continual* fought back to surpass the UrbanPy on Task 4.

Catastrophic Forgetting. To better visualize the forgetting phenomenon on new tasks, we show the training process of our model by using *fine-tune* and *continual* protocols in Figure 3. We have the following findings. For *fine-tune* model, its performances on old tasks are severely degenerated, an obvious consequence of the "catastrophic forgetting". The older the task, the more knowledge it forgets. For *continual* model with our designed AKR algorithm, it successfully alleviates the forgetting problem when learning new tasks. In addition, the *fine-tune* model also faces the overfitting issue on Task-4 at Stage-4 (the validation loss starts to rise), and beyond our expectation, the *continual* model performs surprisingly well. This result suggests another potential benefit of the ARK algorithm.

4.2 Experimental Analysis

Ablation Study. To investigate the contributions of each component in CUFAR, we conduct ablation studies on the following four variants of CUFAR:

- w/o SE: removing the spatial relation extractor.
- w/o TE: removing the temporal feature extractor.

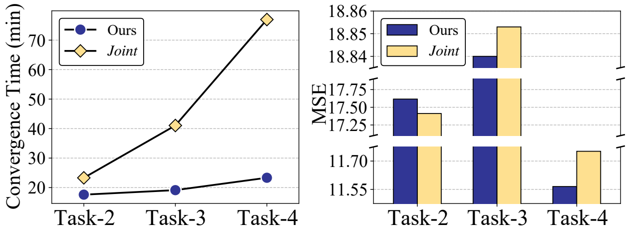

Figure 5: Comparison between *joint* and *continual* (ours) protocols in terms of convergence time and inference error.

- w/o AKR: removing the adaptive knowledge replay, i.e., we use *fine-tune* protocol.
- w/o MMD: removing the maximum mean discrepancy.

Figure 4 shows the ablation results and we have following remarks: (1) The spatial and temporal feature extractors in CUFAR contributes the most. The combination of the two extractors (i.e., the *single-task* model) is superior than either of them alone, demonstrating the effectiveness of the designed inference network. (2) The results of CUFAR along with CUFAR w/o MMD show that the knowledge learned from previous tasks can significantly improve the FUFI performance. Besides, the designed MMD distance in AKR avoids the "negative replaying" issue and enhances the robustness of the algorithm (as the results of Task-4 show).

***Joint* Protocol.** Training all current and previous tasks simultaneously is often considered to be a powerful approach and serves as a soft upper bound of performance (De Lange et al. 2021). However, as we show in Figure 5, *joint* training has two main shortcomings: (1) when there are too many tasks, the computation overhead becomes unaffordable, e.g., *joint* spends 3× more training time than CUFAR on four tasks. (2) noisy data introduced from older tasks may hinder the inference performance if no selection strategy is applied.

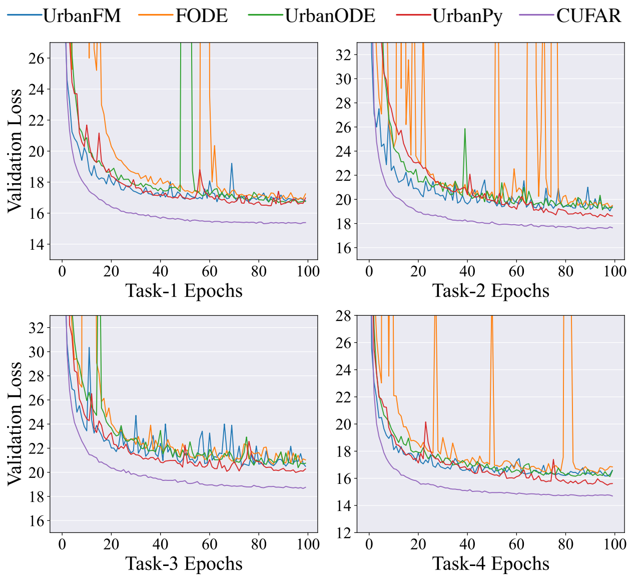

Figure 6: Convergence speeds of CUFAR and baselines.

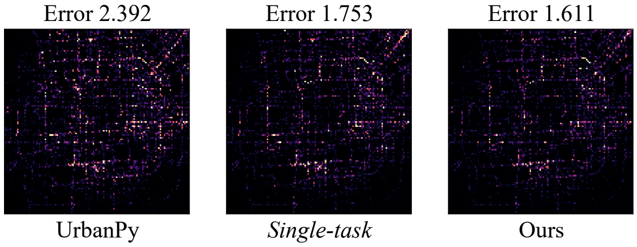

Figure 7: Error visualization.

We can see from the figure that CUFAR equipped with AKR surpasses its *joint* counterpart on Task-3 and Task-4.

Convergence Analysis. Figure 6 shows the validation loss (MSE) during the training phase of CUFAR and baselines on all tasks. Our model converges smoother and faster than baselines while also having the lowest validation losses. For ordinary differential equation (ODE)-based models FODE and UrbanODE, their loss curves oscillated drastically, probably because of the gradient explosion that occurs when solving the ODE functions. UrbanPy, an extension of UrbanFM, employs a cascading strategy that progressively upsamples the coarse-grained flow map, outperforming other baselines. It is worth mentioning that all baselines are less stable on Task-3. This result may be explained by the fact that the flow volumes in TaxiBJ-P3 are larger than in other tasks. Surprisingly, the loss curve of CUFAR keeps steady and smooth as well as on other tasks.

Error Visualization. Figure 7 shows the inference errors $||\widetilde{\mathbf{X}}_{fg} - \mathbf{X}_{fg}||^2$ of three models on a case flow map, the brighter the pixels, the larger the errors. We can observe that the CUFAR has much less brighter pixels than UrbanPy. On certain areas, e.g., the road to Beijing Capital International Airport (the top right corner), CUFAR's inferred fine-grained flow map is more accurate than the *single-task* model.

5 Related Work

Overcoming the catastrophic forgetting in artificial neural networks is attracting numerous research attention (De Lange et al. 2021) due to its significance for a dynamic system that has new data/tasks coming continuously. Continual learning aims to model a sequence of new tasks without forgetting the knowledge of past tasks, which is a promising direction towards sustainable and robust neural networks (Qu et al. 2021; Mai et al. 2022). Early efforts on continual learning can be categorized into three types. (1) *Regularization-based* methods impose restrictions when learning a new task to mitigate catastrophic forgetting. They use specific loss functions to take these constraints on the parameter updating process and consolidate previously learned knowledge (Mai et al. 2022; Li and Hoiem 2017; Kirkpatrick et al. 2017). (2) *Parameter isolation* methods dedicate different model parameters to each task to prevent any possible forgetting (De Lange et al. 2021). When no constraints are applied to the size of the architecture, one can grow new branches for new tasks while freezing the parameters of the old task or provide a model copy to each task (De Lange et al. 2021; Serra et al. 2018; Aljundi, Chakravarty, and Tuytelaars 2017). However, above mentioned two types of methods face several drawbacks. First, they often impose constraints when learning new tasks, which limits the generalizability of the model; Second, they cannot utilize the knowledge from old tasks to improve new task performance. (3) *Replay-based* methods save the data of previous tasks in a memory buffer. When learning new tasks, they replay the samples from the buffer and then mitigate the catastrophic forgetting of previous tasks (Buzzega et al. 2020). Compared to other continual methods, replay-based methods do not constrain the new task optimization to prevent the interference of previous tasks, and they are more suitable for FUFI. Motivated by replay-based methods, we design an adaptive knowledge replay algorithm for selectively replaying the knowledge from previous tasks and finally improving the FUFI performance on new tasks.

6 Conclusions

In this work, we presented CUFAR, a novel continual framework for fine-grained urban flow inference. We propose to utilize the learned knowledge from previous tasks to enhance the learning process of the model on the new task. We designed a spatial-temporal inference network and a general adaptive knowledge replay training algorithm that helps the model alleviate "catastrophic forgetting" and "negative replaying" issues when adapting to new urban flow maps. Extensive experiments on four large-scale real-world FUFI datasets demonstrated the effectiveness and robustness of CUFAR over state-of-the-art baselines.

In our future work, we plan to investigate (1) extending our solution to other urban flow applications/datasets. (2) finding new ways to selectively replay the old samples that are beneficial for the new task with a theoretical guarantee.

7 Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62176043 and No. 62072077), and Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0505).

References

Aljundi, R.; Chakravarty, P.; and Tuytelaars, T. 2017. Expert gate: Lifelong learning with a network of experts. In *CVPR*, 3366–3375.

Arbel, M.; Sutherland, D. J.; Bińkowski, M. a.; and Gretton, A. 2018. On gradient regularizers for MMD GANs. In *NeurIPS*.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *ICLR*.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 15920–15930.

Carpenter, G. A.; and Grossberg, S. 1987. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26(23): 4919–4930.

Chen, G.; Chen, P.; Shi, Y.; Hsieh, C.-Y.; Liao, B.; and Zhang, S. 2019. Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks. *arXiv:1905.05928*.

Cleophas, C.; Cottrill, C.; Ehmke, J. F.; and Tierney, K. 2019. Collaborative urban transportation: Recent advances in theory and practice. *European Journal of Operational Research*, 273(3): 801–816.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 44(7): 3366–3385.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2): 295–307.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *JMLR*, 13(25): 723–773.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 1646–1654.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13): 3521–3526.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 4681–4690.

Li, J.; Wang, S.; Zhang, J.; Miao, H.; Zhang, J.; and Yu, P. 2022. Fine-grained Urban Flow Inference with Incomplete Data. *IEEE TKDE*.

Li, K.; Chen, J.; Yu, B.; Shen, Z.; Li, C.; and He, S. 2020. Supreme: Fine-grained radio map reconstruction via spatial-temporal fusion network. In *IPSN*, 1–12.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE TPAMI*, 40(12): 2935–2947.

Liang, Y.; Ouyang, K.; Jing, L.; Ruan, S.; Liu, Y.; Zhang, J.; Rosenblum, D. S.; and Zheng, Y. 2019. Urbanfm: Inferring fine-grained urban flows. In *SIGKDD*, 3132–3142.

Liang, Y.; Ouyang, K.; Wang, Y.; Liu, X.; Chen, H.; Zhang, J.; Zheng, Y.; and Zimmermann, R. 2022a. TrajFormer: Efficient Trajectory Classification with Transformers. In *CIKM*, 1229—-1237.

Liang, Y.; Ouyang, K.; Wang, Y.; Liu, Y.; Zhang, J.; Zheng, Y.; and Rosenblum, D. S. 2020. Revisiting convolutional neural networks for citywide crowd flow analytics. In *ECML-PKDD*, 578–594.

Liang, Y.; Ouyang, K.; Wang, Y.; Pan, Z.; Yin, Y.; Chen, H.; Zhang, J.; Zheng, Y.; Rosenblum, D. S.; and Zimmermann, R. 2022b. Mixed-Order Relation-Aware Recurrent Neural Networks for Spatio-Temporal Forecasting. *IEEE TKDE*.

Liu, F.; Xu, W.; Lu, J.; Zhang, G.; Gretton, A.; and Sutherland, D. J. 2020. Learning deep kernels for non-parametric two-sample tests. In *ICML*, 6316–6326.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *NIPS*.

Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; and Sanner, S. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469: 28–51.

Ouyang, K.; Liang, Y.; Liu, Y.; Tong, Z.; Ruan, S.; Zheng, Y.; and Rosenblum, D. S. 2022. Fine-Grained Urban Flow Inference. *IEEE TKDE*, 34(06): 2755–2770.

Qu, H.; Rahmani, H.; Xu, L.; Williams, B.; and Liu, J. 2021. Recent advances of continual learning in computer vision: An overview. *arXiv:2109.11369*.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*.

Serra, J.; Suris, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, 4548–4557.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 1874–1883.

Vandal, T.; Kodra, E.; Ganguly, S.; Michaelis, A.; Nemani, R.; and Ganguly, A. R. 2017. DeepSD: Generating high resolution climate change projections through single image super-resolution. In *SIGKDD*, 1663–1672.

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM TOMS*, 11(1): 37–57.

Yu, H.; Xu, X.; Zhong, T.; and Zhou, F. 2022. Fine-grained urban flow inference via normalizing flows (student abstract). In *AAAI*.

Zheng, Y.; Capra, L.; Wolfson, O.; and Yang, H. 2014. Urban computing: concepts, methodologies, and applications. *ACM TIST*, 5(3): 1–55.

Zhong, T.; Yu, H.; Li, R.; Xu, X.; Luo, X.; and Zhou, F. 2022. Probabilistic Fine-Grained Urban Flow Inference with Normalizing Flows. In *ICASSP*, 3663–3667.

Zhou, F.; Jing, X.; Li, L.; and Zhong, T. 2021. Inferring High-Resolutional Urban Flow With Internet Of Mobile Things. In *ICASSP*, 7948–7952.

Zhou, F.; Li, L.; Zhong, T.; Trajcevski, G.; Zhang, K.; and Wang, J. 2020. Enhancing Urban Flow Maps via Neural ODEs. In *IJCAI*, 1295–1302.

Zhou, X.; Zhou, D.; and Liu, L. 2021. TRUFM: a Transformer-Guided Framework for Fine-Grained Urban Flow Inference. In *ICONIP*, 262–273.