Effective Dataset Distillation for Spatio-Temporal Forecasting with Bi-Dimensional Compression

Taehyung Kwon, Yeonje Choi*, Yeongho Kim, Kijung Shin*
KAIST: {taehyung.kwon, yeonjechoi, yeongho, kijungs}@kaist.ac.kr

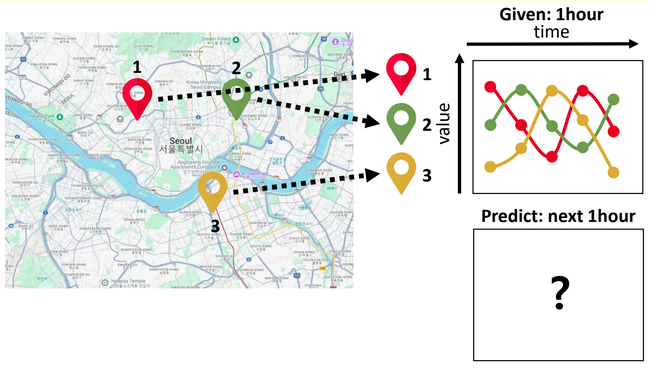
ICDE 2026 MONTRÉAL, CANADA

KAIST

Paper GitHub

Target Application & Research Question

Target Application: Spatio-Temporal Time Series Forecasting



- **Data: Multivariate time series** collected at various locations at regular intervals
- **Given:** All time series data over a certain time period,
- **Task:** To predict the time series for the future time period.

- Spatio-temporal datasets are **massive** in scale, since they are collected from numerous **locations** over extended **periods**.

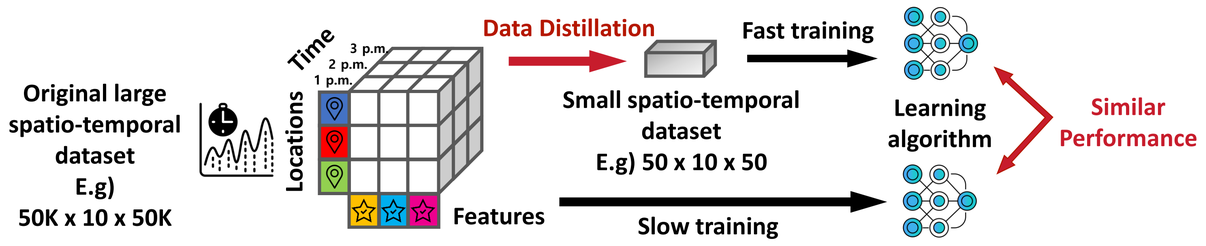
Research Question

"How can we effectively distill a spatio-temporal time series dataset that enables effective and efficient training?"

Problem Definition & Limitation of Existing Methods

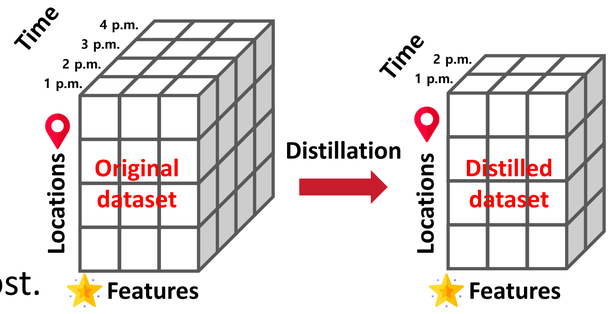
Problem: Spatio-Temporal Time Series Dataset Distillation

- **Goal:** We aim to synthesize a compact spatio-temporal time series dataset that enables effective training in place of the original dataset.



Limitation of Existing Methods

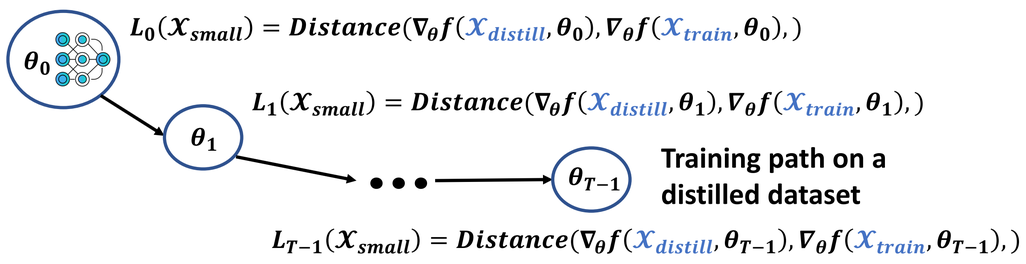
- Simply applying general dataset distillation methods only reduces the number of time steps.
- Distilled datasets still have many locations, leading to high training cost.



Proposed Method: STemDist

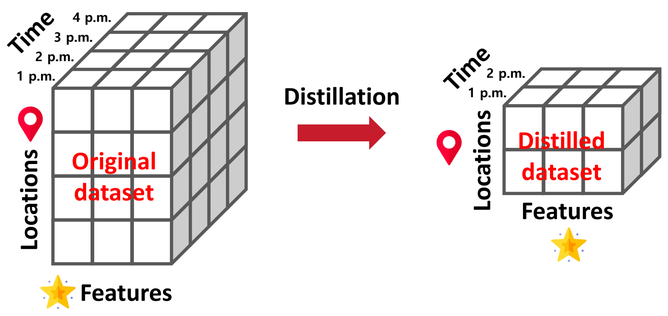
Algorithmic Foundation: Gradient Matching

- **STemDist** is based on gradient matching, which minimizes the distance between the gradients from distilled and original data.



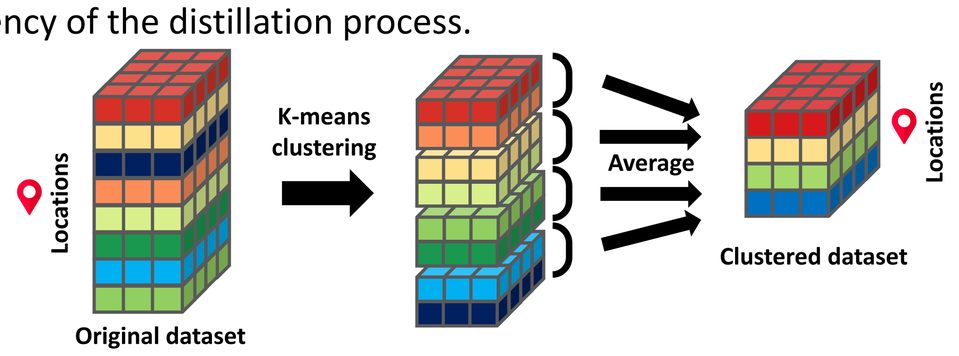
Main Idea: Bi-Dimensional Dataset Compression

- **Motivation:** Naively applying gradient matching reduces only the temporal dimension, while training cost remains high due to many locations.
- **Goal:** To simultaneously compress both temporal and spatial dimensions.
- **Challenge:** Training on a dataset with a smaller number of locations makes inference on the original dataset challenging due to transductive location embeddings in ML models.
- **Our Idea:** To introduce a **shared location encoder** that replaces location embeddings and enables flexible handling of any number of locations.



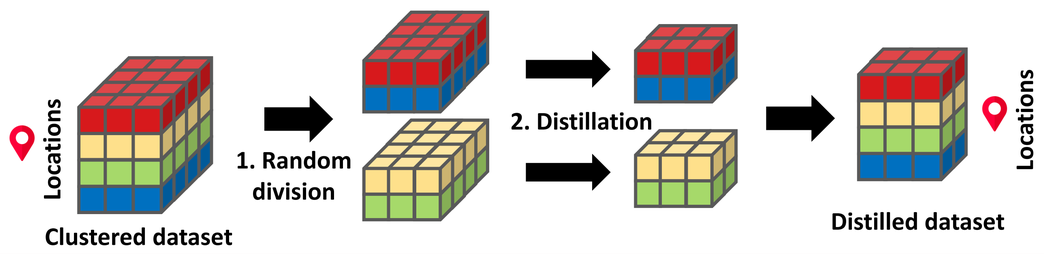
Further Improvement: Location Clustering

- **Motivation:** A large number of locations in the original dataset slows down the distillation process.
- **Our Idea:** To reduce the number of locations in the original dataset via k-means clustering, aligning it with the distilled dataset to improve the efficiency of the distillation process.



Further Improvement: Subset-based Granular Distillation

- **Motivation:** Using all locations together for distillation n cause strong inter-location correlations to dominate, overlooking weak but important ones.
- **Our Idea:** To perform distillation in units of location subsets to allow different parts of the data to be effectively reflected in the distilled dataset.



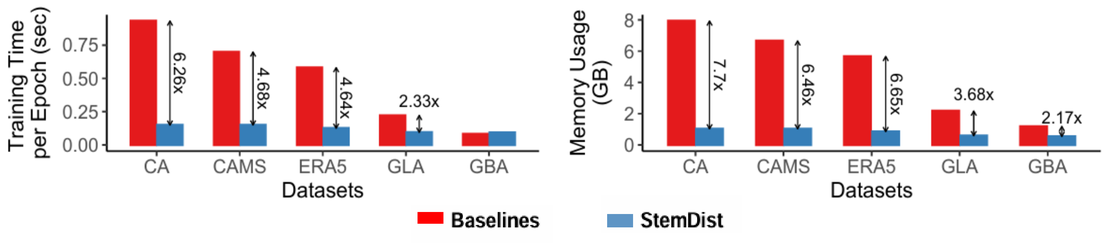
Experiments

Q1. Distillation Performance: ***STemDist is effective.*** Models trained on data distilled by STemDist achieve up to **12% lower prediction error** compared to those trained on data distilled by baselines.

Method	GBA	GLA	ERA5	CA	CAMS
	Relative RMSE	Relative RMSE	Relative RMSE	Relative RMSE	Relative RMSE
Random	0.453 ± 0.086	0.410 ± 0.051	0.520 ± 0.028	0.431 ± 0.065	0.796 ± 0.109
K-Center	0.400 ± 0.042	0.356 ± 0.028	0.524 ± 0.012	0.355 ± 0.034	0.883 ± 0.073
Herding	0.506 ± 0.004	0.428 ± 0.004	0.522 ± 0.001	0.521 ± 0.004	0.730 ± 0.002
CRAIG	0.354 ± 0.007	0.323 ± 0.017	0.470 ± 0.003	0.354 ± 0.021	0.685 ± 0.008
DC	0.348 ± 0.006	0.311 ± 0.009	0.442 ± 0.004	0.331 ± 0.014	0.735 ± 0.006
DM	0.409 ± 0.051	0.380 ± 0.025	0.467 ± 0.008	0.405 ± 0.028	0.674 ± 0.011
MTT	0.414 ± 0.055	0.439 ± 0.066	O.O.M	O.O.M	O.O.M
DATM	0.517 ± 0.007	0.402 ± 0.022	O.O.M	O.O.M	O.O.M
IDM	0.420 ± 0.050	0.381 ± 0.024	0.481 ± 0.018	0.394 ± 0.047	0.709 ± 0.049
Frepo	0.465 ± 0.075	0.439 ± 0.084	0.510 ± 0.031	0.436 ± 0.084	0.802 ± 0.118
CondTSF	0.407 ± 0.041	0.394 ± 0.031	0.489 ± 0.017	0.403 ± 0.051	0.692 ± 0.037
TimeDC	0.422 ± 0.030	0.445 ± 0.068	0.480 ± 0.015	0.448 ± 0.074	0.729 ± 0.078
STemDist	**0.317 ± 0.005**	**0.277 ± 0.002**	**0.426 ± 0.004**	**0.292 ± 0.002**	**0.653 ± 0.005**
Original data	0.264 ± 0.010	0.250 ± 0.005	0.424 ± 0.010	0.253 ± 0.003	0.661 ± 0.004

Compression ratio: 0.5%

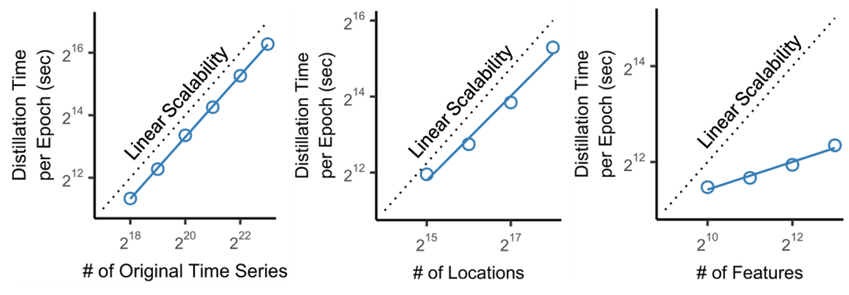
Q2. Efficiency: ***STemDist enables fast and memory-efficient training.*** Model training is up to **6x faster** and requires up to **8x less GPU memory** on data distilled by STemDist than on data distilled by baselines.



Q3. Cross-model Performance: Datasets distilled by **STemDist** are effective for training various models, showing ***strong cross-model generalizability***.

	Graph Wavenet [1]		STGCN [2]		FourierGNN [3]	
	GBA	GLA	GBA	GLA	GBA	GLA
Method	Relative RMSE	Relative RMSE	Relative RMSE	Relative RMSE	Relative RMSE	Relative RMSE
K-Center	0.427 ± 0.032	0.367 ± 0.018	0.714 ± 0.085	0.746 ± 0.062	0.419 ± 0.027	0.395 ± 0.025
DC	0.395 ± 0.033	0.312 ± 0.009	0.812 ± 0.053	0.866 ± 0.031	0.403 ± 0.015	0.376 ± 0.021
CondTSF	0.416 ± 0.066	0.387 ± 0.028	0.697 ± 0.038	0.734 ± 0.087	0.419 ± 0.036	0.421 ± 0.009
STemDist	**0.310 ± 0.007**	**0.273 ± 0.003**	**0.501 ± 0.041**	**0.459 ± 0.034**	**0.384 ± 0.002**	**0.339 ± 0.004**
Original data	0.227 ± 0.009	0.212 ± 0.011	0.294 ± 0.008	0.295 ± 0.003	0.307 ± 0.012	0.284 ± 0.006

Q4. Scalability: ***STemDist is scalable.*** Distillation time of STemDist increases (sub)linearly w.r.t. the number of time series, locations, and features.



Q5. Ablation Study: All key components of **STemDist** help achieve more effective distillation more rapidly.

