Less is More

Efficient Time Series Dataset Condensation via Two-fold Modal Matching

Hao Miao, Ziqiao Lui, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, Christian S. Jensen

February 2025

Proceedings of the VLDB Endowment, Volume 18, Issue 2

Presented by **Andreas Gottschalk Krath**

1. Introduction

IoT

- More devices
- Time series

1.1 Motivation

IoT

- More devices
- Time series

Data Volume

- Too much data
- Edge computing
- What can we do?
  - Less data $\longrightarrow$ Same value

1.1 Motivation

IoT

- More devices
- Time series

Data Volume

- Too much data
- Edge computing
- What can we do?
  - Less data $\longrightarrow$ Same value

Existing Methods

- Coreset construction
- Streaming learning
- Issues
  - Not optimal
  - Downstream tasks

Challenge 1

- Effectiveness
  - Bi-level optimization
  - Capture relevant information
- Generalization
  - Different networks

1.2 Problem

Challenge 1

- Effectiveness
  - Bi-level optimization
  - Capture relevant information
- Generalization
  - Different networks

Solution

- Curriculum Training Trajectory Matching (CT^2D)
- Expert trajectories
  - Based on original dataset
  - Offline
- Match based on model parameters

Challenge 2

- Complex Temporal Dependencies
- Existing methods focus on image
- Time series requires temporal analysis
- Channel independent
  - Training stability

Challenge 2

- Complex Temporal Dependencies
- Existing methods focus on image
- Time series requires temporal analysis
- Channel independent
  - Training stability

Solution

- Time Series Feature Extraction (TSFE)
  - Channel independent mechanism
  - Stacked TSOperators
- Decomposition-Driven Frequency Matching (DDFM)
  - Analyze intermediate TSFE
  - Match to original data

Challenge 3

- Scalability
- Bi-level optimization
- Models must be loaded at runtime
- Ineffective and memory intensive

Challenge 3

- Scalability
- Bi-level optimization
- Models must be loaded at runtime
- Ineffective and memory intensive

Solution

- Expert buffer
  - Pre-computed expert trajectories
- Patching
  - Combine data

Dataset Condensation

- Novel time series dataset condensation
- TSFE
- DDFM
- CT^2M

2. Methodology

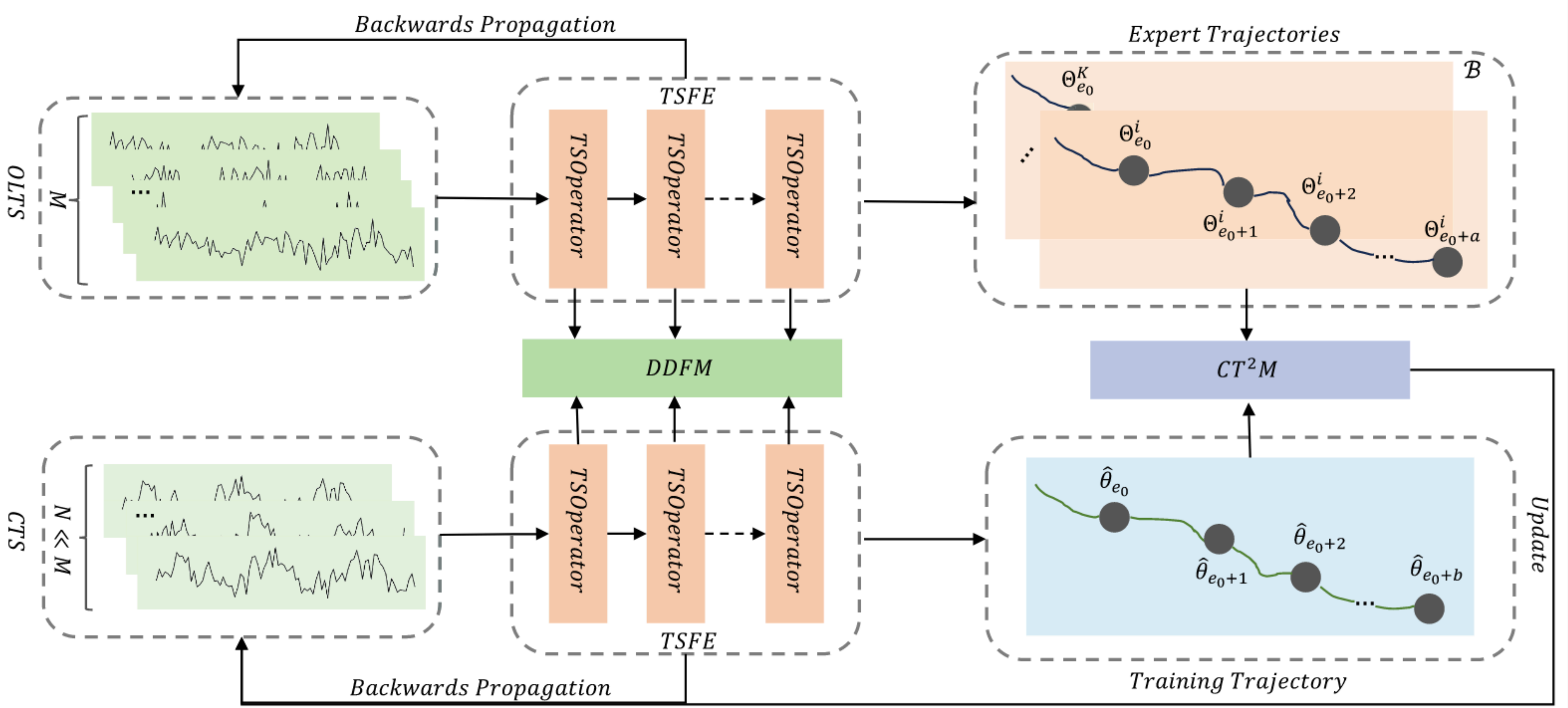


Figure 2: TimeDC Framework Overview

Architecture

- Channel independent mechanism
  - Split variables
- Patching
  - Combine data
- Stacked TSOperators
  - Self-attention
  - Fully connected
  - Normalization
- Concatenate to combine variables

Algorithm 1: Time Series Feature Extraction

Input: a batch of time series: $T_{input} \in \mathcal{B} \times n \times C$; the number of *TSOperators*: N_{op}.
Output: extracted features: h.

$T^1_{input}, \cdots, T^c_{input}, \cdots, T^C_{input} \leftarrow$ Separate T_{input} into C univariate channel-independent time series with Equation 3;
Split each univariate time series into patches;
for $1 \leq c \leq C$ **do**
  $h^0_c \leftarrow T^c_{input}$;
  for $0 < j < N_{op}$ **do**
    Feature extraction with Equation 4;
    $h^j_c \leftarrow TSOperator(h^{j-1}_c)$;
  $h_c \leftarrow h^{N_{op}-1}_c$;
$h \leftarrow$ Concatenate $\{h_c\}^C_{c=1}$ along the channel;
return h.

$$h^j = TSOperator(h^{j-1}) = Norm(FC(MultiHead(h^{j-1}))), \quad (4)$$

Architecture

- Works on intermediate TSFE output
- Splits representation
  - Trends (long-term)
  - Seasonality (the rest)
- Consine similarity to match
- Produces a “data-related” loss term
  - Relevant for back propagation

$$h^j_{TRE} = AvgPool(Padding(h^j)),\ h^j_{SEA} = h^j - h^j_{TRE}, \tag{6}$$

$$\begin{aligned} cos(h^j_{TRE_\mathcal{T}}, h^j_{TRE_\mathcal{S}}) &= \frac{h^j_{TRE_\mathcal{T}}}{||h^j_{TRE_\mathcal{T}}||_2} \cdot \frac{h^j_{TRE_\mathcal{S}}}{||h^j_{TRE_\mathcal{S}}||_2} \\ cos(h^j_{SEA_\mathcal{T}}, h^j_{SEA_\mathcal{S}}) &= \frac{h^j_{SEA_\mathcal{T}}}{||h^j_{SEA_\mathcal{T}}||_2} \cdot \frac{h^j_{SEA_\mathcal{S}}}{||h^j_{SEA_\mathcal{S}}||_2}, \end{aligned} \tag{8}$$

2.4 Curriculum Training Trajectory Matching

Architecture

- Expert buffer and current model
  - Current model is based on expert
- Train current model a epochs
- Compare cosine for parameter movement
- Rank based on similarity
  - Easy $\longrightarrow$ Hard
- Compute trajectory matching loss

Algorithm 2: Curriculum Training Trajectory Query and Matching

Input: A buffer $\mathcal{B}$ with a set of trajectories pre-trained on the original TS dataset $\mathcal{T}$ parameterized by $\{\Theta_{\mathcal{T}}^{k}\}_{k=1}^{K}$; current model parameters $\widetilde{\theta}^{\mathcal{S}}$ on $\mathcal{S}$.

Output: Trajectory matching loss L_{tmm}.

Distance list $DT \leftarrow []$;
Pre-update $\widetilde{\theta}^{\mathcal{S}}$ for a-steps with Equation 13;
$\widetilde{\theta}_{e_0+a} \leftarrow \widetilde{\theta}_{e_0} - \sum_{s=1}^{a}(\alpha\nabla\mathcal{L}(f_{\theta^{\mathcal{S}}}, \mathcal{S}))$;
for $\Theta_{\mathcal{T}}^{k} \in \{\Theta_{\mathcal{T}}^{k}\}_{k=1}^{K}$ **do**
    Compute the distance dis_k between $\widetilde{\theta}|_{e_0}^{a}$ and $\Theta_{\mathcal{T}}^{k}$ with Equation 15;
    $dis_k \leftarrow -D(\widetilde{\theta}|_{e=e_0}^{a}, \theta^{k}|_{e=e_0}^{a})$;
    $DT \leftarrow (k, dis_k)$;
Rank DT in a descending order;
$\beta \leftarrow 0$
while $\beta < K$ **do**
    $k \leftarrow DT[\beta][0]$
    $L_{tmm} \leftarrow$ Sample trajectory $\Theta_{\mathcal{T}}^{k}$ and match the training trajectory according to Equation 11;
    $\beta \leftarrow \beta + 1$;
return L_{tmm}

Loss

- Task specific loss
  - How good is the data?
- Frequency related loss
  - How similar is the data?
- Trajectory matching loss
  - How does data affect training?

$$L_{\mathrm{all}} = \mathcal{L} + L_{\mathrm{Fre}} + L_{\mathrm{tmm}} \quad (16)$$

3. Experiments

Table 1: Overall Performance Comparison on Seven Datasets

Baseline		Random		K-Center		Herding		DC		MTT		TimeDC		Whole Dataset	
Dataset	*PL*	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Weather	96	0.731	1.256	0.452	0.687	0.478	0.677	0.361	0.514	0.295	0.244	**0.257**	**0.188**	*0.239*	*0.182*
	192	0.786	1.302	0.487	0.723	0.512	0.688	0.413	0.527	0.344	0.301	**0.285**	**0.247**	*0.261*	*0.195*
	336	0.794	1.311	0.524	0.756	0.554	0.712	0.444	0.567	0.368	0.328	**0.330**	**0.287**	*0.282*	*0.241*
Traffic	96	0.675	1.125	0.503	0.576	0.483	0.554	0.375	0.603	0.279	0.403	**0.254**	**0.375**	*0.247*	*0.337*
	192	0.712	1.144	0.514	0.604	0.517	0.606	0.432	0.633	0.336	0.442	**0.297**	**0.405**	*0.265*	*0.338*
	336	0.729	1.117	0.523	0.611	0.553	0.654	0.449	0.676	0.355	0.471	**0.312**	**0.423**	*0.297*	*0.360*
Electricity	96	0.421	0.669	0.448	0.583	0.501	0.592	0.376	0.513	0.296	0.283	**0.274**	**0.267**	*0.252*	*0.268*
	192	0.450	0.743	0.476	0.601	0.534	0.628	0.419	0.532	0.315	0.337	**0.285**	**0.294**	*0.239*	*0.255*
	336	0.491	0.853	0.506	0.622	0.569	0.477	0.436	0.544	0.339	0.356	**0.304**	**0.322**	*0.271*	*0.285*
ETTh1	96	0.523	0.745	0.554	0.698	0.536	0.656	0.503	0.442	0.456	0.464	**0.413**	**0.401**	*0.354*	*0.386*
	192	0.557	0.786	0.578	0.722	0.589	0.698	0.552	0.508	0.504	0.471	**0.436**	**0.428**	*0.362*	*0.355*
	336	0.588	0.802	0.604	0.745	0.603	0.723	0.556	0.513	0.498	0.464	**0.447**	**0.431**	*0.409*	*0.387*
ETTh2	96	0.487	0.655	0.589	0.711	0.521	0.589	0.463	0.524	0.388	0.342	**0.368**	**0.271**	*0.324*	*0.255*
	192	0.509	0.673	0.605	0.732	0.553	0.621	0.488	0.536	0.416	0.384	**0.389**	**0.302**	*0.332*	*0.257*
	336	0.524	0.689	0.623	0.744	0.564	0.640	0.505	0.540	0.435	0.455	**0.411**	**0.334**	*0.376*	*0.296*
ETTm1	96	0.743	1.124	0.525	0.492	0.607	0.554	0.603	0.665	0.512	0.453	**0.503**	**0.442**	*0.453*	*0.403*
	192	0.764	1.245	0.566	0.510	0.628	0.571	0.597	0.647	0.563	0.501	**0.512**	**0.465**	*0.464*	*0.432*
	336	0.801	1.128	0.571	0.523	0.644	0.582	0.624	0.668	0.552	0.488	**0.500**	**0.483**	*0.477*	*0.455*
ETTm2	96	0.664	0.795	0.486	0.623	0.524	0.558	0.472	0.535	0.376	0.421	**0.354**	**0.391**	*0.347*	*0.381*
	192	0.687	0.804	0.512	0.643	0.549	0.583	0.488	0.567	0.453	0.479	**0.401**	**0.421**	*0.358*	*0.403*
	336	0.702	0.823	0.558	0.661	0.598	0.624	0.493	0.556	0.473	0.523	**0.453**	**0.474**	*0.406*	*0.435*

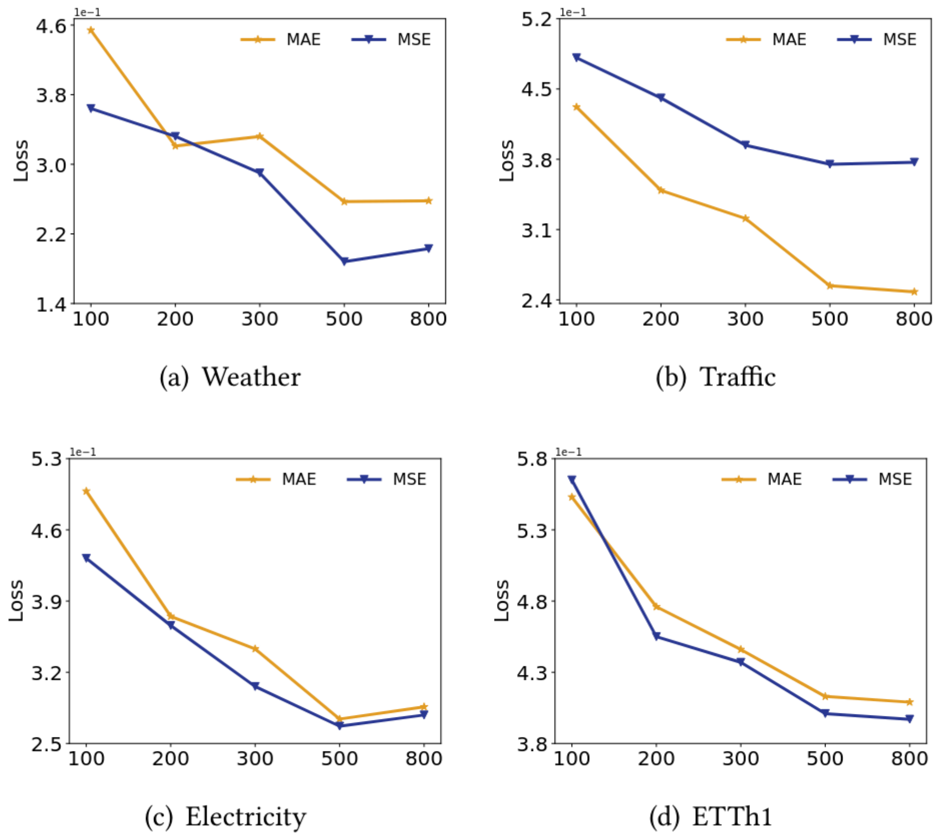


Figure 5: Effect of the Size of Condensed TS Dataset on Four Datasets ($PL = 96$)

3.3 Ablation Study

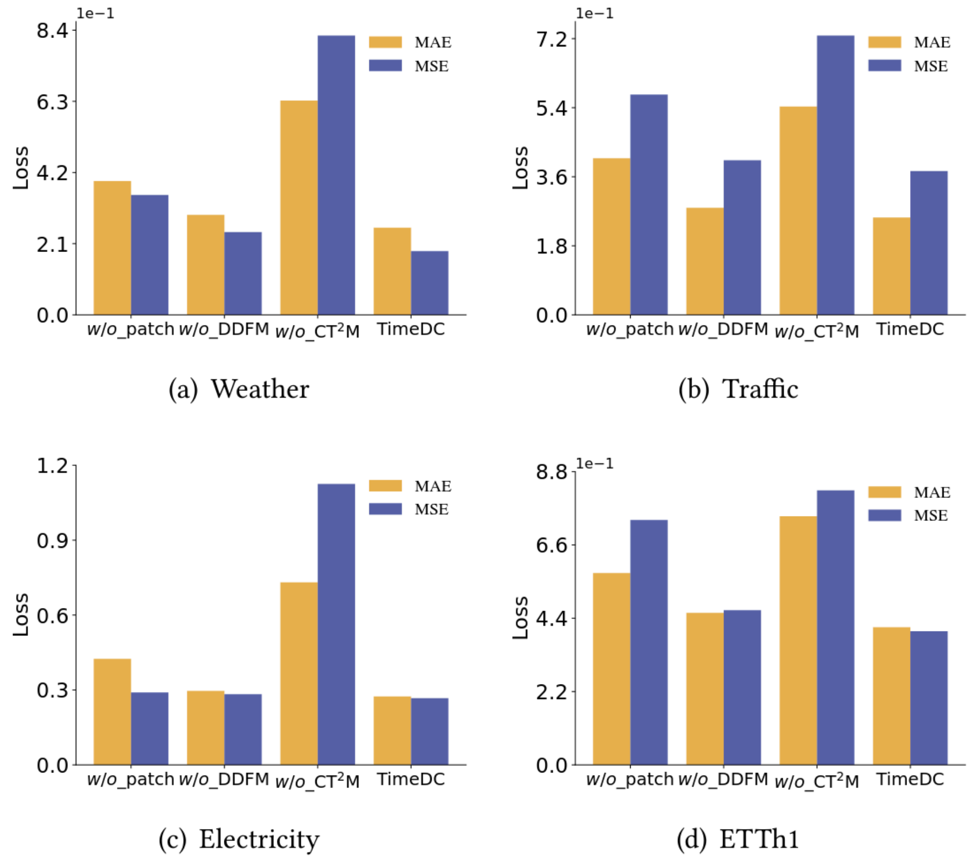


Figure 6: Performance of TimeDC and Its Variants on Four Datasets (PL = 96)

Table 2: Cross-Architecture Performance Comparison

Method	Metric	PL	Weather	Traffic	Electricity	ETT			
						ETTh1	ETTh2	ETTm1	ETTm2
TimeDC	MAE	96	**0.257**	**0.254**	**0.274**	**0.413**	**0.368**	**0.503**	**0.354**
		192	**0.285**	**0.297**	**0.285**	**0.436**	**0.389**	**0.512**	**0.401**
	MSE	96	**0.188**	**0.375**	**0.267**	**0.401**	**0.271**	**0.442**	**0.391**
		192	**0.247**	**0.405**	**0.294**	**0.428**	**0.302**	**0.465**	**0.421**
Autoformer	MAE	96	<u>0.312</u>	<u>0.370</u>	<u>0.343</u>	<u>0.453</u>	<u>0.473</u>	<u>0.548</u>	<u>0.342</u>
		192	<u>0.381</u>	<u>0.385</u>	<u>0.355</u>	<u>0.478</u>	<u>0.491</u>	<u>0.550</u>	<u>0.334</u>
	MSE	96	<u>0.255</u>	0.597	<u>0.236</u>	<u>0.465</u>	<u>0.412</u>	<u>0.542</u>	<u>0.265</u>
		192	<u>0.334</u>	0.613	<u>0.264</u>	<u>0.493</u>	<u>0.488</u>	<u>0.532</u>	<u>0.287</u>
Informer	MAE	96	0.423	0.430	0.428	0.773	0.842	0.576	0.552
		192	0.482	0.476	0.446	0.788	0.954	0.597	0.532
	MSE	96	0.354	0.643	0.253	0.992	1.032	0.624	0.402
		192	0.478	0.710	0.271	0.987	1.055	0.653	0.432
Transformer	MAE	96	0.389	0.412	0.398	0.632	0.506	0.563	0.555
		192	0.588	0.431	0.422	0.612	0.513	0.588	0.576
	MSE	96	0.344	<u>0.578</u>	0.267	0.785	0.579	0.615	0.479
		192	0.524	<u>0.567</u>	0.288	0.732	0.542	0.643	0.455

3.5 Efficiency

Table 3: Dynamic Tensor Memory Cost on Four Datasets

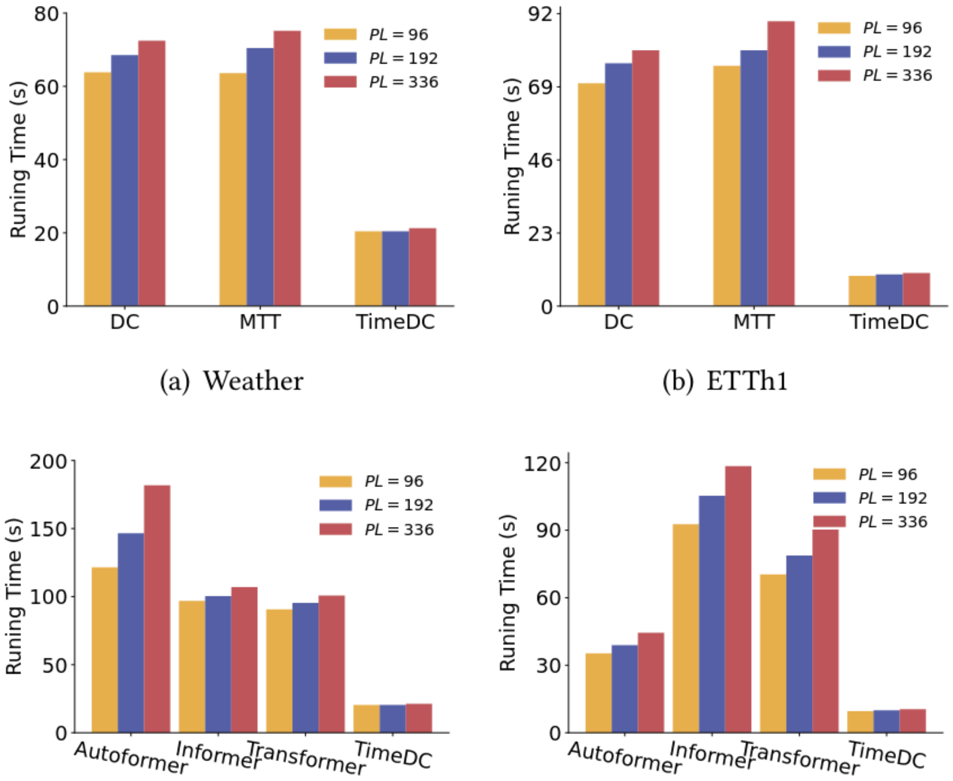
Dataset	DC	MTT	TimeDC
Weather	10.0 GB	8.9 GB	3.3 GB
Traffic	17.8 GB	13.7 GB	10.9 GB
Electricity	8.5 GB	932.5 MB	516.0 MB
ETTh1	1.9 GB	845.5 MB	280.9 MB

Table 4: Training Time of TimeDC and Training Time on Condensed and Original Datasets (s/epoch)

Dataset	TimeDC	Condensed Dataset	Original Dataset
Weather	22.39	4.31	35.26
Traffic	232.34	61.94	346.76
Electricity	314.56	41.14	522.85
ETTh1	14.38	4.93	20.43

Table 6: Storage Comparison on Four Datasets

Storage	Weather	Traffic	Electricity	ETTh1
Whole Dataset	2.9 GB	20.0 GB	11.2 GB	313.1 MB
Condensed TS	38.5 MB	827.5 MB	308.2 MB	12.8 MB



(a) Weather (b) ETTh1

(c) Weather (d) ETTh1

Figure 7: Training Time Comparison

4. Critique

TimeDC

- Condensation framework
- TSFE prediction architecture
- Interchangable
  - Understand from context

Table 1: Overall Performance Comparison on Seven Datasets

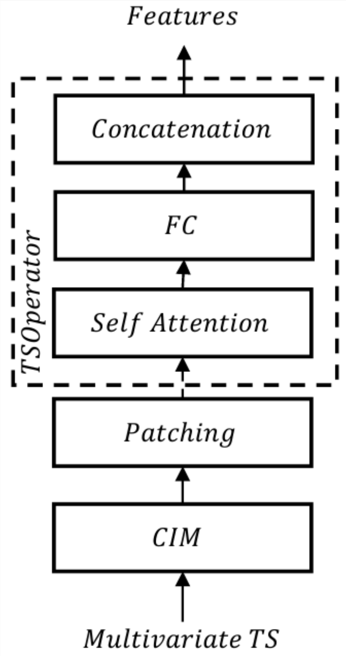
Baseline		Random		K-Center		Herding		DC		MTT		TimeDC		Whole Dataset	
Dataset	*PL*	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Weather	96	0.731	1.256	0.452	0.687	0.478	0.677	0.361	0.514	0.295	0.244	**0.257**	**0.188**	*0.239*	*0.182*
	192	0.786	1.302	0.487	0.723	0.512	0.688	0.413	0.527	0.344	0.301	**0.285**	**0.247**	*0.261*	*0.195*
	336	0.794	1.311	0.524	0.756	0.554	0.712	0.444	0.567	0.368	0.328	**0.330**	**0.287**	*0.282*	*0.241*
Traffic	96	0.675	1.125	0.503	0.576	0.483	0.554	0.375	0.603	0.279	0.403	**0.254**	**0.375**	*0.247*	*0.337*
	192	0.712	1.144	0.514	0.604	0.517	0.606	0.432	0.633	0.336	0.442	**0.297**	**0.405**	*0.265*	*0.338*
	336	0.729	1.117	0.523	0.611	0.553	0.654	0.449	0.676	0.355	0.471	**0.312**	**0.423**	*0.297*	*0.360*
Electricity	96	0.421	0.669	0.448	0.583	0.501	0.592	0.376	0.513	0.296	0.283	**0.274**	**0.267**	*0.252*	*0.268*
	192	0.450	0.743	0.476	0.601	0.534	0.628	0.419	0.532	0.315	0.337	**0.285**	**0.294**	*0.239*	*0.255*
	336	0.491	0.853	0.506	0.622	0.569	0.477	0.436	0.544	0.339	0.356	**0.304**	**0.322**	*0.271*	*0.285*
ETTh1	96	0.523	0.745	0.554	0.698	0.536	0.656	0.503	0.442	0.456	0.464	**0.413**	**0.401**	*0.354*	*0.386*
	192	0.557	0.786	0.578	0.722	0.589	0.698	0.552	0.508	0.504	0.471	**0.436**	**0.428**	*0.362*	*0.355*
	336	0.588	0.802	0.604	0.745	0.603	0.723	0.556	0.513	0.498	0.464	**0.447**	**0.431**	*0.409*	*0.387*
ETTh2	96	0.487	0.655	0.589	0.711	0.521	0.589	0.463	0.524	0.388	0.342	**0.368**	**0.271**	*0.324*	*0.255*
	192	0.509	0.673	0.605	0.732	0.553	0.621	0.488	0.536	0.416	0.384	**0.389**	**0.302**	*0.332*	*0.257*
	336	0.524	0.689	0.623	0.744	0.564	0.640	0.505	0.540	0.435	0.455	**0.411**	**0.334**	*0.376*	*0.296*
ETTm1	96	0.743	1.124	0.525	0.492	0.607	0.554	0.603	0.665	0.512	0.453	**0.503**	**0.442**	*0.453*	*0.403*
	192	0.764	1.245	0.566	0.510	0.628	0.571	0.597	0.647	0.563	0.501	**0.512**	**0.465**	*0.464*	*0.432*
	336	0.801	1.128	0.571	0.523	0.644	0.582	0.624	0.668	0.552	0.488	**0.500**	**0.483**	*0.477*	*0.455*
ETTm2	96	0.664	0.795	0.486	0.623	0.524	0.558	0.472	0.535	0.376	0.421	**0.354**	**0.391**	*0.347*	*0.381*
	192	0.687	0.804	0.512	0.643	0.549	0.583	0.488	0.567	0.453	0.479	**0.401**	**0.421**	*0.358*	*0.403*
	336	0.702	0.823	0.558	0.661	0.598	0.624	0.493	0.556	0.473	0.523	**0.453**	**0.474**	*0.406*	*0.435*

Table 2: Cross-Architecture Performance Comparison

Method	Metric	PL	Weather	Traffic	Electricity	ETT			
						ETTh1	ETTh2	ETTm1	ETTm2
TimeDC	MAE	96	**0.257**	**0.254**	**0.274**	**0.413**	**0.368**	**0.503**	**0.354**
		192	**0.285**	**0.297**	**0.285**	**0.436**	**0.389**	**0.512**	**0.401**
	MSE	96	**0.188**	**0.375**	**0.267**	**0.401**	**0.271**	**0.442**	**0.391**
		192	**0.247**	**0.405**	**0.294**	**0.428**	**0.302**	**0.465**	**0.421**
Autoformer	MAE	96	0.312	0.370	0.343	0.453	0.473	0.548	0.342
		192	0.381	0.385	0.355	0.478	0.491	0.550	0.334
	MSE	96	0.255	0.597	0.236	0.465	0.412	0.542	0.265
		192	0.334	0.613	0.264	0.493	0.488	0.532	0.287
Informer	MAE	96	0.423	0.430	0.428	0.773	0.842	0.576	0.552
		192	0.482	0.476	0.446	0.788	0.954	0.597	0.532
	MSE	96	0.354	0.643	0.253	0.992	1.032	0.624	0.402
		192	0.478	0.710	0.271	0.987	1.055	0.653	0.432
Transformer	MAE	96	0.389	0.412	0.398	0.632	0.506	0.563	0.555
		192	0.588	0.431	0.422	0.612	0.513	0.588	0.576
	MSE	96	0.344	0.578	0.267	0.785	0.579	0.615	0.479
		192	0.524	0.567	0.288	0.732	0.542	0.643	0.455

TSOperator Structure

- Figure 3
  - Self-attention
  - Fully connected
  - Concatenation
- Equation 4 + Algorithm 1
  - Self-attention
  - Fully connected
  - Normalization



$$h^j = TSOperator(h^{j-1}) = Norm(FC(MultiHead(h^{j-1}))), \quad (4)$$

DDFM

- No details on comparison
  - Full condensed dataset?
  - Single time series from original?
  - Average similarity?

4.2 Method

CT²M

- Unclear how loss works
  - Accumulated?
  - Sequential backpropagation?
- Why order similarity?

Algorithm 2: Curriculum Training Trajectory Query and Matching

Input: A buffer $\mathcal{B}$ with a set of trajectories pre-trained on the original TS dataset $\mathcal{T}$ parameterized by $\{\Theta^k_{\mathcal{T}}\}^K_{k=1}$; current model parameters $\widetilde{\theta}^{\mathcal{S}}$ on $\mathcal{S}$.

Output: Trajectory matching loss L_{tmm}.

Distance list $DT \leftarrow []$;

Pre-update $\widetilde{\theta}^{\mathcal{S}}$ for a-steps with Equation 13;

$\widetilde{\theta}_{e_0+a} \leftarrow \widetilde{\theta}_{e_0} - \sum_{s=1}^{a}(\alpha\nabla\mathcal{L}(f_{\theta^S}, \mathcal{S}))$;

for $\Theta^k_{\mathcal{T}} \in \{\Theta^k_{\mathcal{T}}\}^K_{k=1}$ **do**

  Compute the distance dis_k between $\widetilde{\theta}|^a_{e_0}$ and $\Theta^k_{\mathcal{T}}$ with Equation 15;

  $dis_k \leftarrow -D(\widetilde{\theta}|^a_{e=e_0}, \theta^k|^a_{e=e_0})$;

  $DT \leftarrow (k, dis_k)$;

Rank DT in a descending order;

$\beta \leftarrow 0$

while $\beta < K$ **do**

  $k \leftarrow DT[\beta][0]$

  $L_{tmm} \leftarrow$ Sample trajectory $\Theta^k_{\mathcal{T}}$ and match the training trajectory according to Equation 11;

  $\beta \leftarrow \beta + 1$;

return L_{tmm}

Dataset Condensation Performance (Table 1)

- All datasets
- 96, 192, 336 PL

Generalization Performance (Table 2)

- All datasets
- 96, 192 PL
  - No explination for 336 missing

Dynamic Memory Cost (Table 3)

- Only 4 datasets
- Only ETTh1 as sub-dataset

Condensed Dataset Size (Figure 5)

- Only 4 datasets
- Only ETTh1 as sub-dataset
- Only 96 PL
  - No explination
- Same for ablation study (Figure 6)

Training Time for Dataset Condensation

- Only 4 datasets (same as others)
- 96, 192, 336 PL again

5. Project & Course

Current Project

- Kind of opposites
  - We have little data, would like more
- Explore how “dense” our data is
  - New usecase for TimeDC

Current Project

- Kind of opposites
  - We have little data, would like more
- Explore how "dense" our data is
  - New usecase for TimeDC

Previous Projects

- Trajectory simplification
- Not synthesized data
- Also focused on use case
  - Query-driven trajectory simplification

Course

- Memformer
- Three shared authors
- Time series forecasting (long-term)
- Patching
- Channel independence

6. Evaluation

Idea

- Logical methodology
  - Data loss (DDFM)
  - Training loss (CT²M)
  - Usecase loss (downstream prediction task)
- Focus on efficiency
  - Faster training
  - Less data
  - Lower memory usage

6. Evaluation

Execution

- Inconsistencies
  - TSOperator structure
  - TimeDC terminology
- Missing details
  - Loss in CT^2M
- Cherry-picked results
  - Datasets
  - Forecasting horizons
  - Bold in Table 2
- Resource constrained environments
  - Edge computing?
  - Embedded?
  - Moved computation to offline