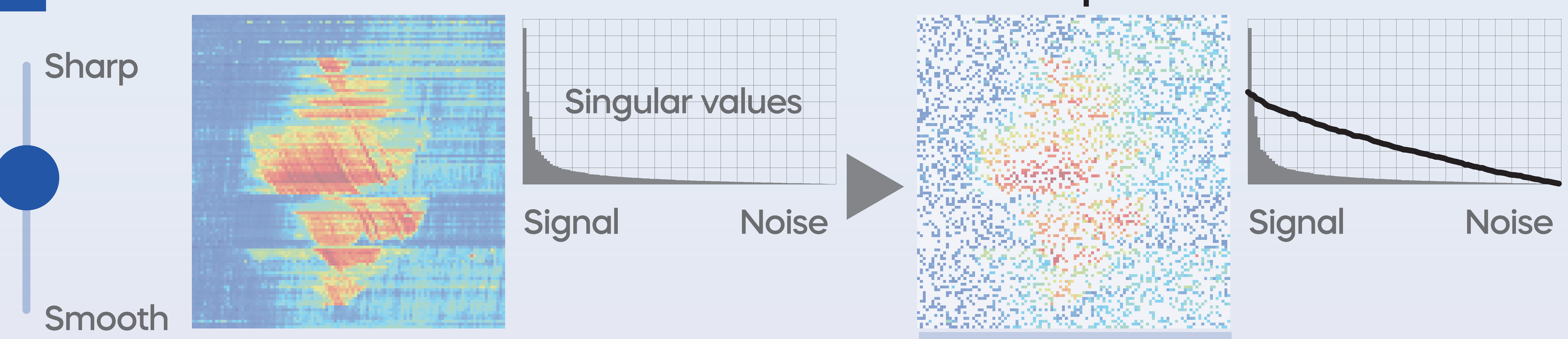
0 ¡ADVERTENCIA! THE PREVALENT SPATIOTEMPORAL MISSING DATA PROBLEM



IN STATISTICS, MISSING DATA OCCURS WHEN NO VALUE IS STORED FOR A VARIABLE IN AN OBSERVATION

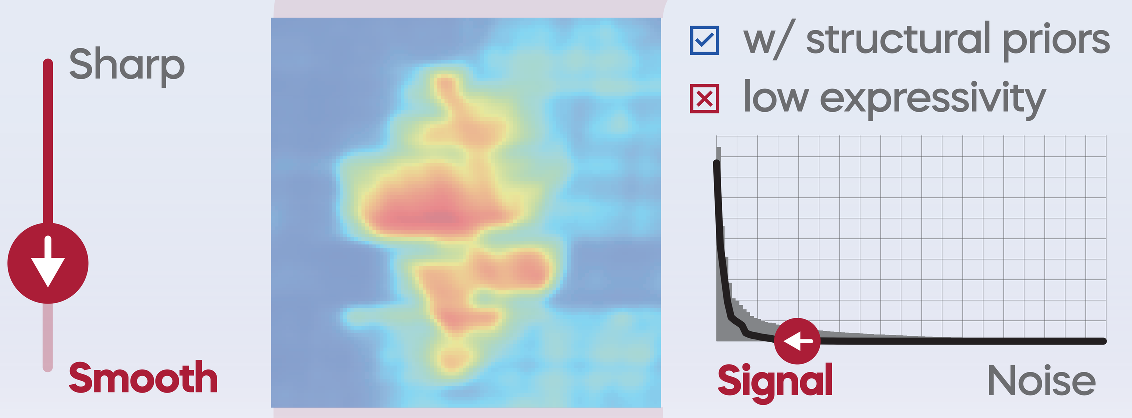
- Missing data is common in detection systems, especially high-resolution monitoring: e.g. transportation, energy, and environmental systems
- Prevalent in real-world scientific and engineering problems

FACTORS LIKE WEATHER, ENERGY SUPPLY, AND SENSOR MAINTENANCE CAN IMPACT DATA QUALITY

- Urban Air project: ~30% invalid records due to station malfunction
- Uber Movement data (e.g., New York City): ~85% missing records

1 "THE IMPUTER'S DILEMMA" SPATIOTEMPORAL IMPUTATION IS OFTEN PRONE TO DEVIATION FROM GROUND TRUTH, EITHER BEING OVERLY SMOOTH AND OVERLY SHARP

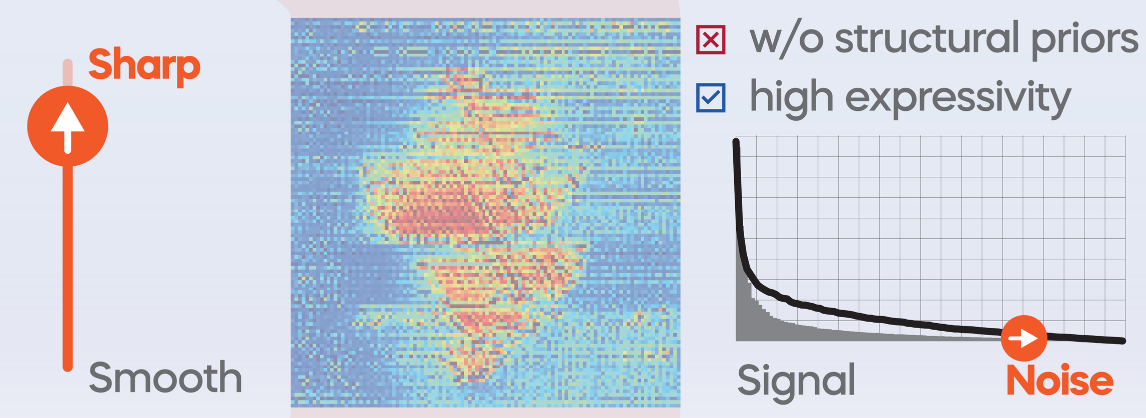
PARADIGM 1
Low-rank/ low-dim imputation



PROS
Effective priors for spatiotemporal data

CONS
Struggles with complex patterns; over-smoothes reconstructions, losing informative signals

PARADIGM 2
Deep learning imputation

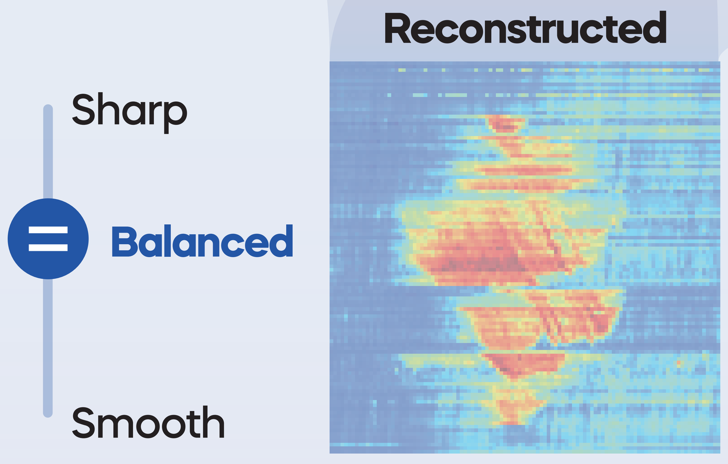


PROS
High expressivity and performance

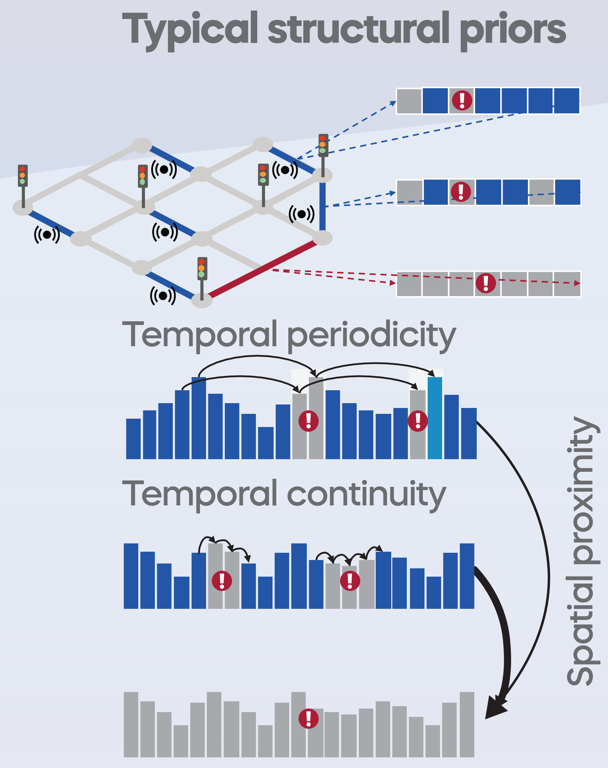
CONS
Data-intensive, requiring substantial training; prone to overfitting due to lack of structural priors

2 INTRODUCING IMPUTEFORMER, A REMEDY CAREFULLY DESIGNED TO REGULARIZE MODEL EXPRESSIVITY W/ STRUCTURAL PRIORS, STRIKING A BALANCE BETWEEN SMOOTH AND SHARP IMPUTATIONS

PARADIGM 1x2
Low rankness-induced deep learning



- w/ structural priors
- high expressivity (+model priors)



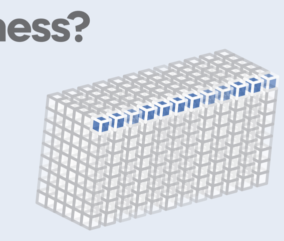
Architectural Overview

1. $\mathcal{Z}^{(0)}_{t:t+T} = \texttt{InputEmb}(\mathbf{X}_{t:t+T}, \mathbf{U}_{t:t+T}, \mathbf{V})$,
2. $\mathcal{Z}^{(\ell+1)}_{t:t+T} = \texttt{TemporalInteraction}(\mathcal{Z}^{(\ell)}_{t:t+T})$,
3. $\mathcal{Z}^{(\ell+1)}_{t:t+T} = \texttt{SpatialInteraction}(\mathcal{Z}^{(\ell+1)}_{t:t+T}), \forall \ell \in \{0, \dots L\}$,
4. $\widehat{\mathbf{X}}_{t:t+T} = \texttt{Readout}(\mathcal{Z}^{(L+1)}_{t:t+T})$.

How do we incorporate the structural priors w/ low-rankness?

TEMPORAL PROJECTED ATTENTION

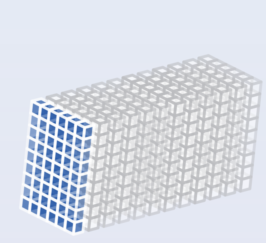
- Temporal patterns governed by low-dim structures
- In self-attention, we incorporate low-rank priors with a shared learnable vector to project temporal dimensions to a lower space



$$
\begin{aligned}
\widetilde{\mathbf{Z}}^{i,(\ell)}_{\text{proj}} &= \texttt{SelfAtten}(\mathbf{P}^{(\ell)}_{\text{proj}}, \mathbf{Z}^{i,(\ell)}, \mathbf{Z}^{i,(\ell)})),\\
&= \texttt{Softmax}\left(\frac{\mathbf{P}^{(\ell)}_{\text{proj}}\mathbf{W}_{\text{Q}}\mathbf{W}_{\text{K}}^{\top}\mathbf{Z}^{i,(\ell),\top}}{\sqrt{D'}}\right)\mathbf{Z}^{i,(\ell)}\mathbf{W}_{\text{V}}\\
\mathbf{Z}^{i,(\ell)}_{\text{hat}} &= \texttt{SelfAtten}(\mathbf{Z}^{i,(\ell)}, \mathbf{P}^{(\ell)}_{\text{proj}}, \widetilde{\mathbf{Z}}^{i,(\ell)}_{\text{proj}})),\\
&= \texttt{Softmax}\left(\frac{\mathbf{Z}^{i,(\ell)}\mathbf{W}_{\text{Q}}\mathbf{W}_{\text{K}}^{\top}\mathbf{P}^{(\ell),\top}_{\text{proj}}}{\sqrt{D'}}\right)\widetilde{\mathbf{Z}}^{i,(\ell)}_{\text{proj}}\mathbf{W}_{\text{V}}\\
\widehat{\mathbf{Z}} &= \texttt{SelfAtten}(\mathbf{Q}, \mathbf{P}, \texttt{SelfAtten}(\mathbf{P}, \mathbf{K}, \mathbf{V})),\\
&= \sigma(\mathbf{Q}\mathbf{P}^{\top})\texttt{SelfAtten}(\mathbf{P}, \mathbf{K}, \mathbf{V}),\\
&= \sigma(\mathbf{Q}\mathbf{P}^{\top})\sigma(\mathbf{P}\mathbf{K}^{\top})\mathbf{V} \approx \frac{1}{N^2}\mathbf{Q}(\mathbf{P}^{\top}\mathbf{P})\mathbf{K}^{\top}\mathbf{V}
\end{aligned}
$$

SPATIAL EMBEDDED ATTENTION

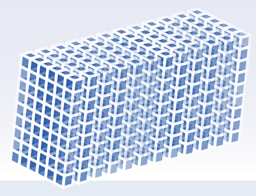
- Applying Transformer blocks to spatial dimensions poses challenges
- We address this by using node embeddings as compact representations



$$
\begin{aligned}
\mathbf{Q}^{(\ell)}_e &= \texttt{Linear}(\mathbf{E}),\ \mathbf{K}^{(\ell)}_e = \texttt{Linear}(\mathbf{E}),\\
\mathbf{A}^{(\ell)} &= \texttt{Softmax}\left(\frac{\mathbf{Q}^{(\ell)}_e\mathbf{K}^{(\ell)\top}_e}{\sqrt{D'}}\right)\\
\mathbf{A}^{(\ell)} &\approx \sigma_2(\widetilde{\mathbf{Q}}^{(\ell)}_e)\sigma_1(\widetilde{\mathbf{K}}^{(\ell)}_e)^{\top},\\
\texttt{SelfAtten}(\mathbf{E}, \mathbf{E}, \mathbf{Z}) &= \sigma(\mathbf{E}\mathbf{W}_E)\sigma(\mathbf{E}\mathbf{W}_E)^{\top}\mathbf{Z}\mathbf{W}_V,\\
&\quad r \leq \min\{N, D_{\text{emb}}\}
\end{aligned}
$$

FOURIER SPECTRUM LOSS

- We constrain the estimated tensor's rank
- However, directly optimizing tensor rank is challenging due to non-trivial, non-differentiable computations
- Discrete Fourier Transform is used to ease this process

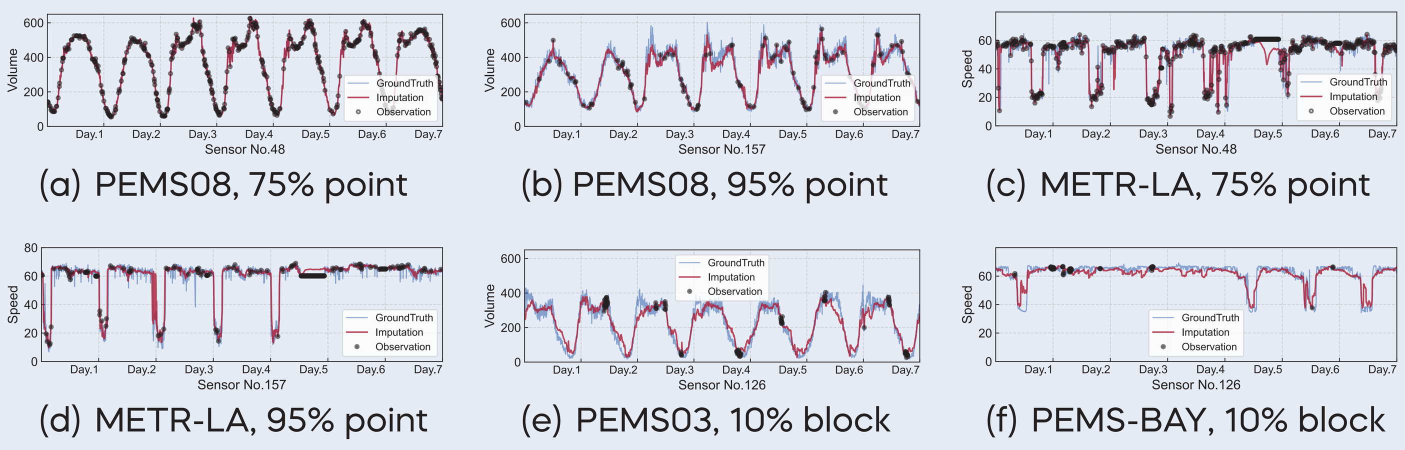


$$
\begin{aligned}
\bar{\mathbf{X}} &= \mathbf{M}_{\text{missing}} \odot \widehat{\mathbf{X}} + (1 - \mathbf{M}_{\text{missing}}) \odot \mathbf{Y},\\
\mathcal{L}_{\text{FIL}} &= \frac{1}{NT}\sum \|\texttt{Flatten}(\texttt{FFT}(\bar{\mathbf{X}}, \text{dim} = [0, 1]))\|_1
\end{aligned}
$$

3 WELL-PERFORMED IN GENERAL SURPASSING EXISTING BASELINES IN EXTENSIVE EXPERIMENTS ACROSS DIVERSE SPATIOTEMPORAL DATASETS, MISSING RATES, AND MISSING PATTERNS

	Point missing						Block missing					
Models	PEMS-BAY	METR-LA	PEMS03	PEMS04	PEMS07	PEMS08	PEMS-BAY	METR-LA	PEMS03	PEMS04	PEMS07	PEMS08
Average	5.45	7.52	85.30	103.61	122.35	89.51	5.48	7.43	85.56	103.82	123.05	89.42
MICE	2.82	2.89	20.07	28.60	37.11	30.26	2.36	2.73	21.90	32.45	37.20	26.66
TRMF	2.10	3.51	18.80	24.34	29.06	20.27	2.09	3.36	18.71	24.47	29.42	19.80
LRTC-AR	0.94	2.14	15.52	22.11	27.60	19.33	4.05	5.35	17.59	24.08	27.82	19.95
Bi-MPGRU	0.72	2.00	11.23	15.84	15.66	11.90	1.41	2.33	13.87	19.81	21.12	15.89
rGAIN	1.90	2.81	13.32	22.86	24.41	16.33	2.21	2.95	14.85	23.26	26.69	27.12
BRITS	1.84	2.42	12.74	20.00	23.97	15.78	1.91	2.40	12.93	19.80	23.26	16.37
SAITS	1.33	2.25	12.40	20.23	22.81	15.12	1.58	2.32	12.43	20.35	22.82	16.80
Transformer	0.76	2.18	12.04	16.76	16.86	12.58	1.69	3.58	24.07	29.63	33.14	25.61
ST-Transformer	0.75	2.19	11.44	16.22	15.84	12.10	1.71	3.58	23.55	29.17	32.14	24.67
TIDER	1.43	2.68	15.02	22.17	21.38	18.46	2.46	4.95	21.12	23.74	28.66	21.00
TimesNet	1.47	2.93	14.99	20.40	22.00	16.53	2.73	4.79	44.85	51.05	60.90	45.78
GRIN	0.68	1.91	10.31	16.25	11.90	12.33	1.20	2.08	12.28	23.23	16.04	19.69
SPIN	0.79	1.93	12.85	18.96	17.61	15.02	1.13	2.02	14.68	19.85	16.99	16.81
ImputeFormer	**0.64** 5.9% ↓	**1.80** 5.8% ↓	**8.23** 20.2% ↓	**14.92** 5.8% ↓	**11.38** 4.4% ↓	**11.01** 7.5% ↓	**0.95** 15.9% ↓	**1.86** 7.9% ↓	**9.02** 26.5% ↓	**16.83** 15.0% ↓	**13.82** 13.8% ↓	**12.50** 21.3% ↓

	SOLAR		CER-EN		Simulated faults	
Models	Point missing	Block missing	Point missing	Block missing	AQI 36	AQI
Average	7.60	7.56	0.583	0.596	61.81	43.78
MICE	1.59	1.58	0.535	0.555	38.90	29.12
TRMF	2.44	2.35	0.557	0.559	41.91	27.67
Bi-MPGRU	N.A.	N.A.	0.247	0.349	12.02	15.41
rGAIN	1.52	1.64	0.418	0.440	15.69	22.13
BRITS	1.28	1.34	0.351	0.366	14.74	20.72
SAITS	0.98	1.25	0.341	0.368	19.79	21.09
Transformer	2.19	3.58	0.254	0.353	14.99	17.04
ST-Transformer	2.17	3.57	0.251	0.351	13.27	18.55
TIDER	2.84	3.87	0.336	0.377	32.85	18.11
TimesNet	2.93	4.73	0.328	0.460	32.30	28.99
GRIN	N.A.	N.A.	**0.235**	0.341	12.08	14.51
SPIN	N.A.	N.A.	OOM	OOM	11.89	14.31
ImputeFormer	**0.51** 48.0% ↓	**0.89** 28.8% ↓	0.236 0.4% ↑	**0.296** 13.2% ↓	**11.58** 2.6% ↓	**13.40** 6.4% ↓



(a) PEMS08, 75% point (b) PEMS08, 95% point (c) METR-LA, 75% point

(d) METR-LA, 95% point (e) PEMS03, 10% block (f) PEMS-BAY, 10% block

INTERPRETATIONS

ON TEMPORAL PROJECTOR

- Temporal modes reconstructing neural representations for imputation, similar to low-rank reconstruction
- Post-projection, hidden states have lower SVs than incomplete inputs and resemble the complete representations

ON SPATIAL EMBEDDING

- Incomplete data yield noisy correlations
- Learned attention maps approximate actual ones
- Node embedding shows low-rank distribution, serving as a dense sensor space surrogate

ON FOURIER SPECTRUM

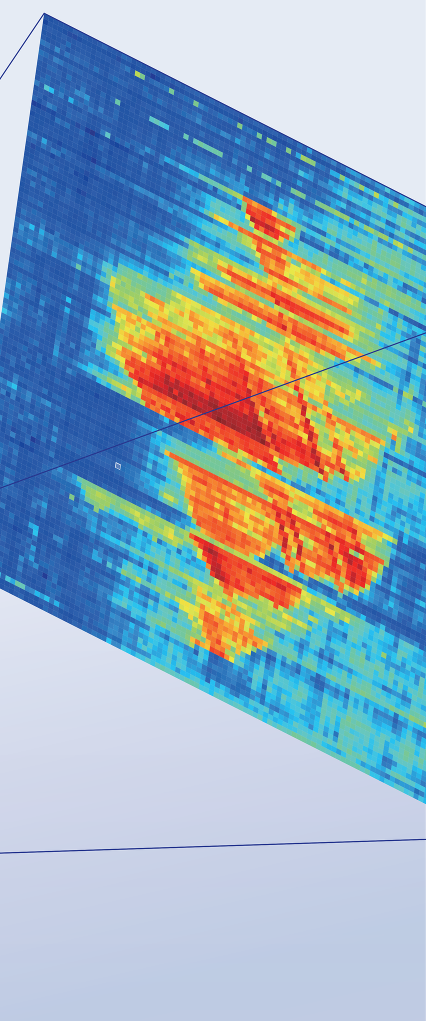
- Transformers downplay dominant SVs w/o explicit low-rank modeling;
- Pure low-rank models like MF oversmooth by constraining energy to the spectrum's first part
- ImputeFormer balances signals and high-frequency noise effectively

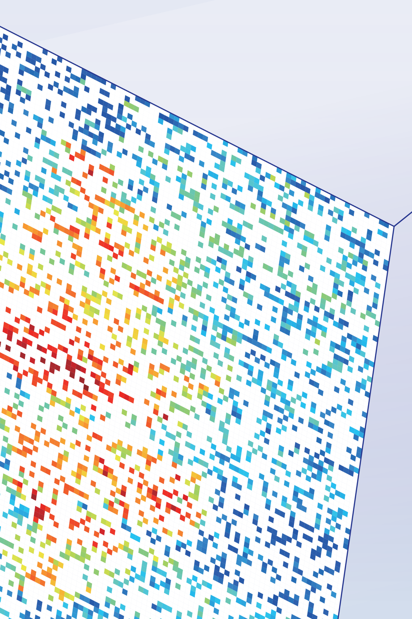
HIGHLIGHTS

- ImputeFormer: Low-rank Imputation Transformers
- Novel structural priors: projected temporal & embedded spatial attention
- New regularization: Fourier spectrum sparsity loss
- Extensive spatiotemporal data experiments

TAKEAWAYS

- Time series primitives enhance data-driven deep models
- Potential for general data imputation architectures
- Representation learning and multipurpose pretraining for time series





同濟大學 TONGJI UNIVERSITY
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

KDD2024
BARCELONA, ESPAÑA

SCAN TO VIEW OUR FULL PAPER

Tong Nie[1,2] Guoyang Qin[1] Wei Ma[2] Yuewen Mei[1] Jian Sun[1]

LOW RANKÑESS-INDUCED TRANSFORMERS
FOR GENERALIZABLE SPATIOTEMPORAL IMPUTATION