STaRFormer: Semi-Supervised Task-Informed Representation Learning via Dynamic Attention-Based Regional Masking for Sequential Data

M. Forstenhäusler[1, 2], D. Külzer[1], C. Anagnostopoulos[2], S. Parambath[2] and N. Weber[1]

[1]BMW Group, [2]University of Glasgow

Motivation

- Intent prediction from user-trajectory recorded with a Digital Key (DK) in the near vicinity of the vehicle.

- General assumptions for sequential modelling [1]:
  - fully observed
  - stationary
  - sampled at regular intervals

- In real-world scenarios → these assumptions often do not apply [2].

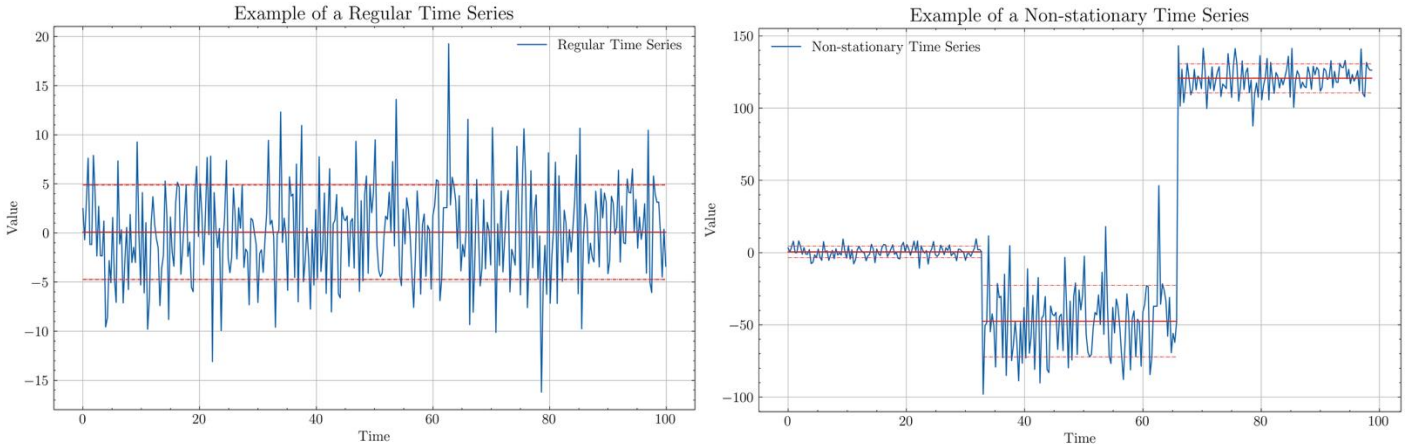
- Digital Keys Trajectory (DKT) data collection results in:
  - Non-stationary sequential data
    - ~79% is non-stationary, confirmed by KPSS and ADF tests.
  - Irregularly sampled sequential data,
    - Ultra-Wideband (UWB) ranging collects measurements at irregularly sampled time intervals.

BMW Digital Key

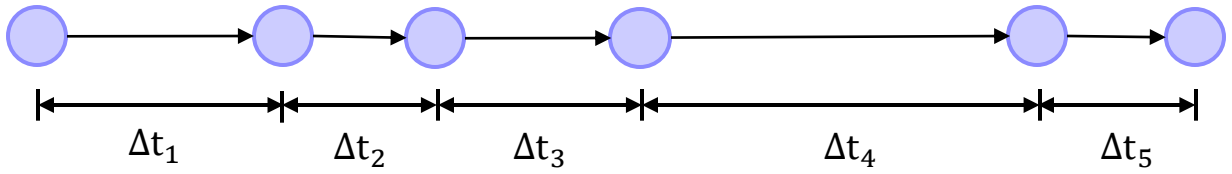
Use-Case Demo

[1] Zekun Li, Shiyang Li, and Xifeng Yan. Time Series as Images: Vision Transformer for Irregularly Sampled Time Series. *Advances in Neural Information Processing Systems,* 36:49187–49204, December 2023.
[2] Luke Bermingham and Ickjai Lee. A probabilistic stop and move classifier for noisy GPS trajectories. Data Min. Knowl. Discov., 32(6):1634–1662, November 2018. ISSN 1384-5810. doi: 10.1007/s10618-018-0568-8.

Motivation & Approach



Regular time series vs non-stationary time series.



Example of an irregular sampled time series.

Goal

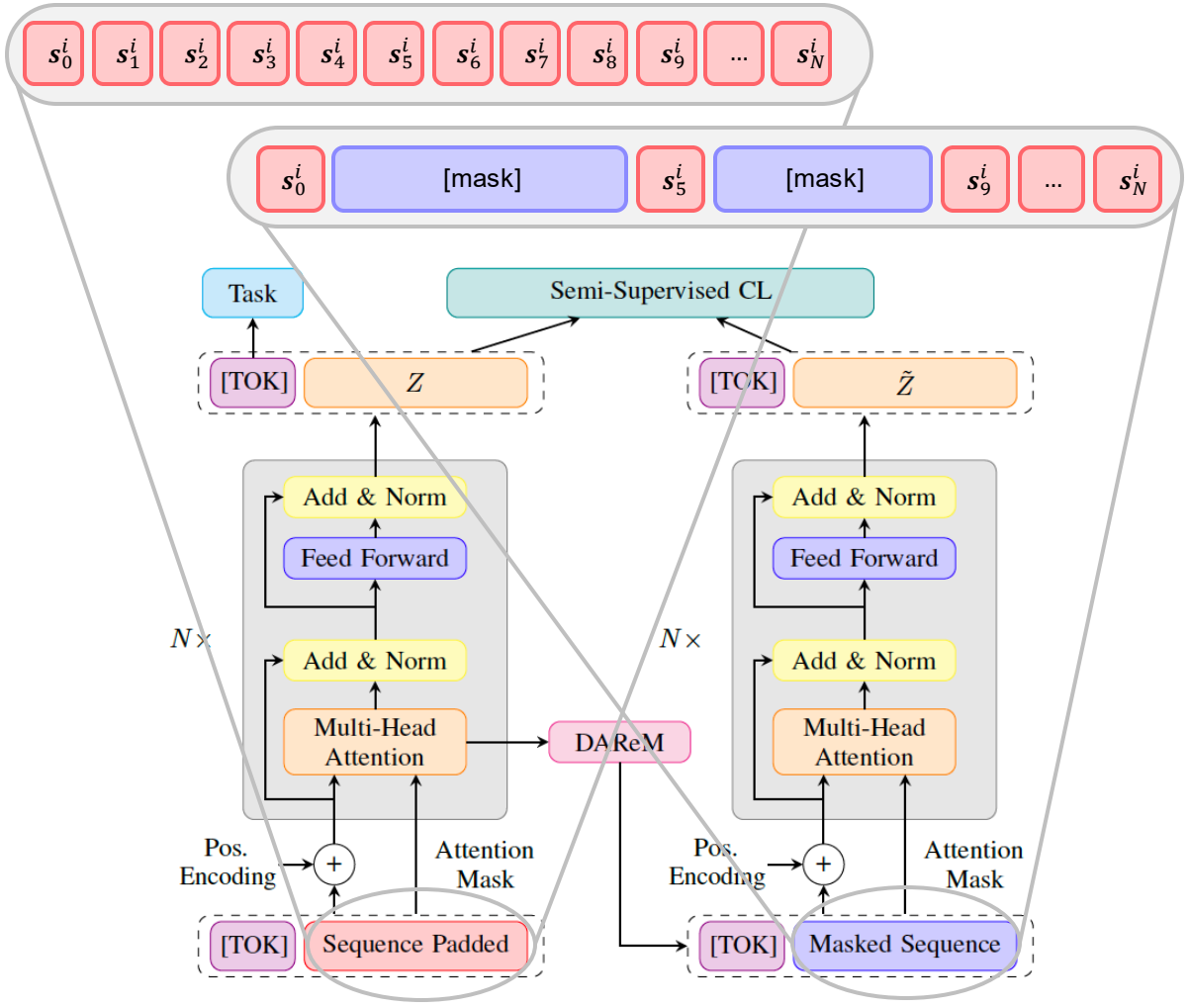
- Robust modeling technique for irregularities in time series.
- Supports binary classification (BMW Use-Case) but can easily be extended to other tasks.

Approach

- Create robust latent embeddings to handle irregularities in sequential data.
- Enhance model latent space → improve the downstream task performance.

STaRFormer

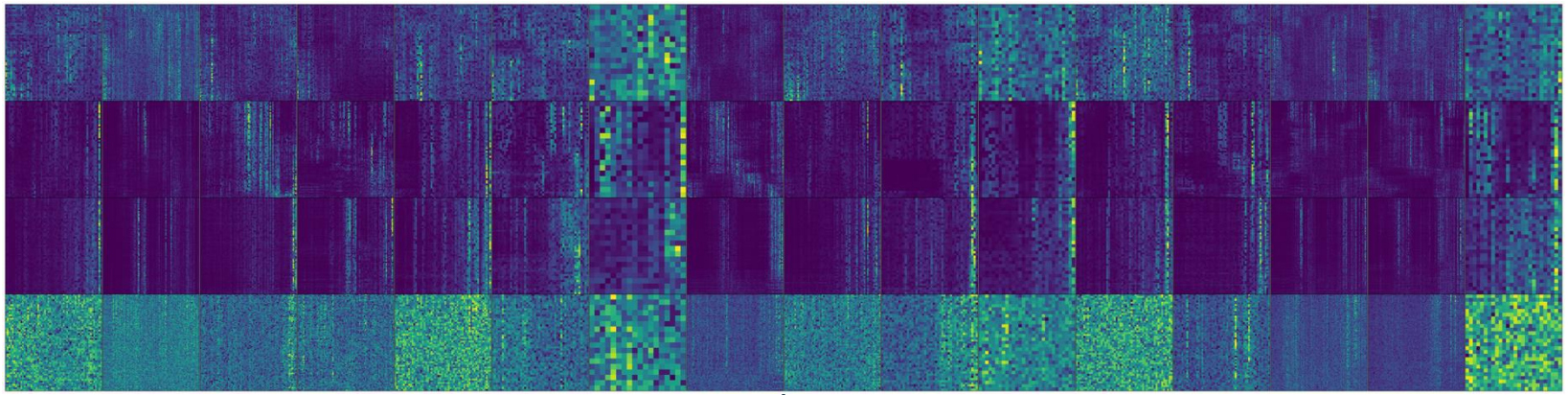
1. Encoder-only Transformer as backbone to extract features from sequential data.
2. Employ **D**ynamic **A**ttention-Based **Re**gional **M**asking **(DAReM)** to manipulate important task-specific regions within an input sequence.
    1. Introduces synthetic variations in statistical properties (non-stationary).
    2. Simulates varying sampling frequencies (irregular sampling).
3. Employ a novel semi-supervised CL approach which maximizes agreement between **batch-wise** and **class-wise similarities** in the latent space → creates robust task-informed latent representations.
    - Incorporation of **DAReM** during training process creates two correlated latent representations → ***masked*** and ***unmasked***.



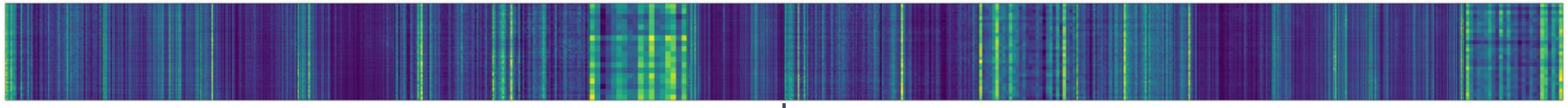
STaRFormer architecture.

STaRFormer – **D**ynamic **A**ttention-Based **Re**gional **M**asking **(DAReM)**

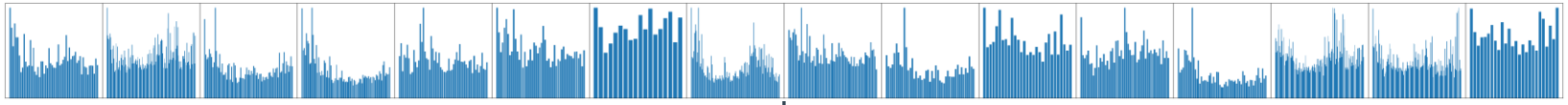
Gathered attention weights per batch (→) for each layer (↓) of the transformer.



Aggregation via attention rollout.



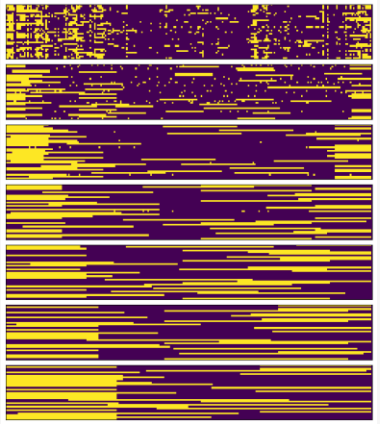
Compute attention scores per element.



Creation of regional mask around most important elements.

Creation of regional mask

- φ: determines the total number of elements to mask.
- ξ: determines the number of top-k elements to mask based on the attention scores.
- γ: determines the bounds of the region to mask.



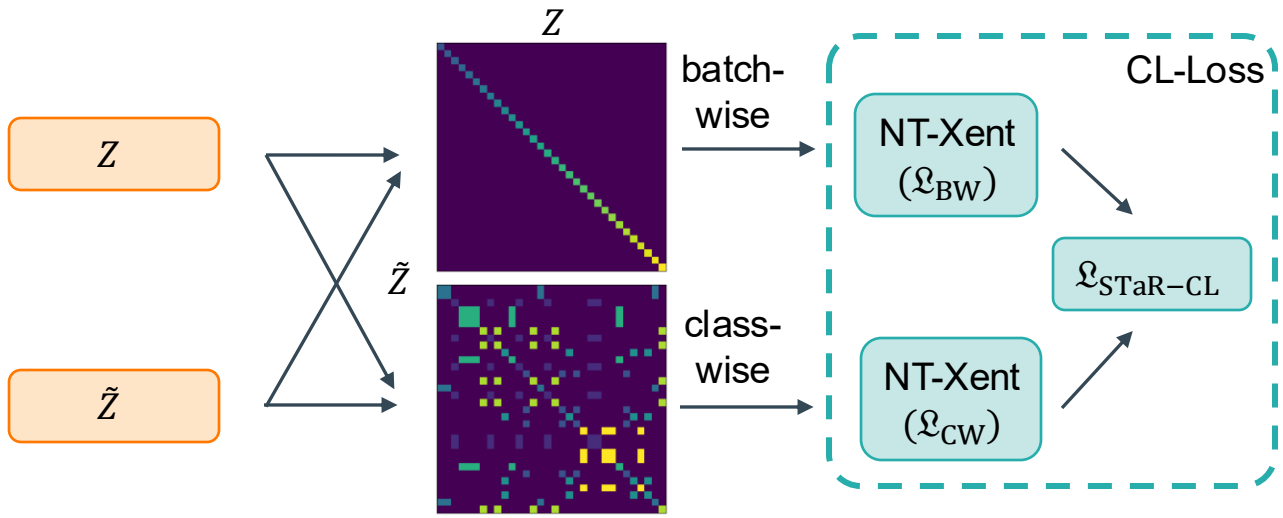
Regional masks with different region parameter γ.

STaRFormer – Semi-Supervised Contrastive Learning (CL)

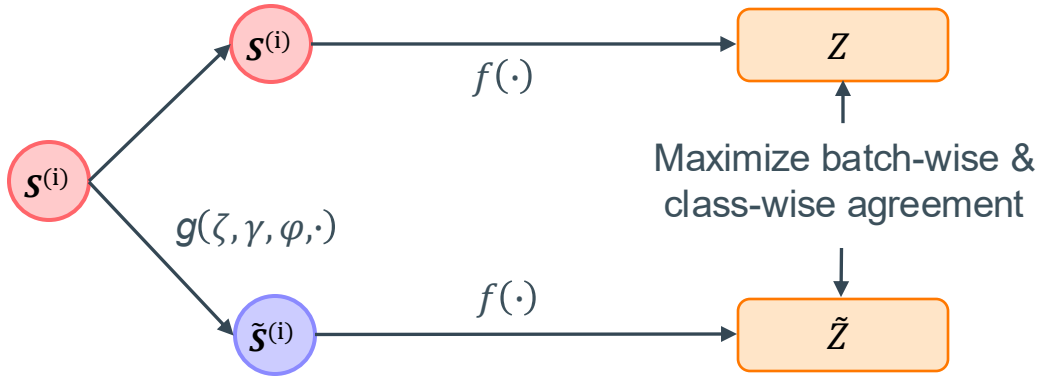
- Via **DAReM,** we obtain two correlated latent space representations, $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Z}}$.

- Facilitate semi-supervised CL by leveraging:
  - Batch-wise positive pairs: masked ($\tilde{\boldsymbol{Z}}^{(i)}$) and unmasked ($\boldsymbol{Z}^{(i)}$) embeddings of the same input sequence $\boldsymbol{S}^{(i)}$ (**self-supervised**).
  - Class-wise positive pairs: masked ($\tilde{\boldsymbol{Z}}$) and unmasked embeddings ($\boldsymbol{Z}$) of the same class (**supervised**).

- Goal: Maximize agreement between masked, $\tilde{\boldsymbol{Z}}$, and unmasked, $\boldsymbol{Z}$, latent space representation.



Semi-supervised contrastive loss formulation



Semi-supervised CL approach

Experiments

1. Classification

A. Spatiotemporal, non-stationary, and irregularly sampled benchmark (custom)
   - Consists of 2 datasets: DKT and a public real-world dataset from Microsoft.

B. Irregularly sampled benchmark
   - Consists of 3 public datasets.

C. Regular time series benchmark (UEA)
   - Consists of 30 public datasets.

2. Anomaly Detection

- Evaluation on a public benchmark.
- Consists of 2 datasets.
- Compare against 6 Models.

3. Regression

- Evaluation on a public benchmark.
- Consists of 19 datasets.
- Compare against 15 models.

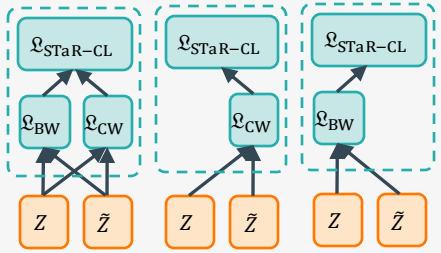
Ablation Studies

1. Architecture

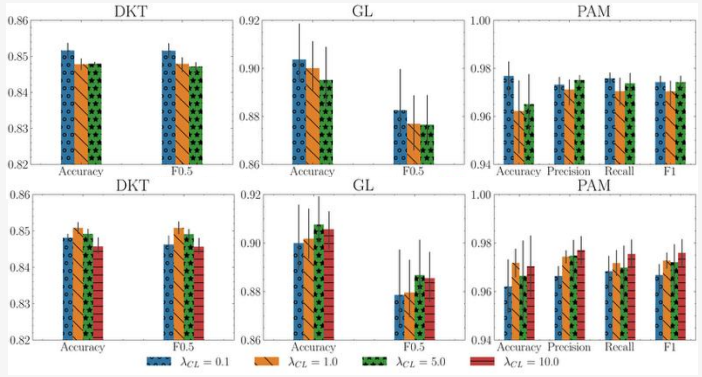
- Comparison of STaRFormer against 2 ablations (Transformer-only & random regional masking) on 19 datasets → STaRFormer outperforms both ablations.

2. Impact of semi-supervised CL approach

- Fusing batch-wise and class-wise similarities is beneficial (semi-supervised).
- Emphasis on CL loss, i.e., higher λ_{CL}, improves performance.



(a) Semi-supervised (b) w/o self-supervised (c) w/o supervised

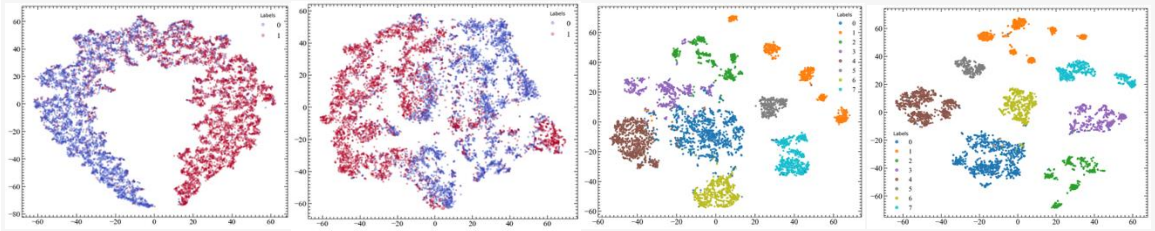


Results semi-supervised CL ablation

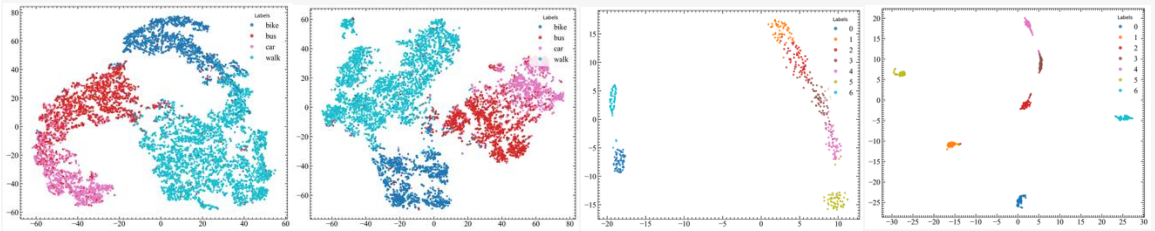
3. Impact of the regional masking approach

- **No** top performance for masking individual elements ($\gamma = 0$) → masking regions improves performance.

4. Latent space analysis



(a) DKT, Base (b) DKT, STaRFormer (c) PAM, Base (d) PAM, STaRFormer



(e) GL, Base (f) GL, STaRFormer (g) PAM, Base (h) PAM, STaRFormer

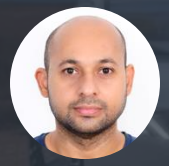
Comparison of sample latent spaces of Base (Transformer-only) and STaRFormer, visualized using t-SNE.

NEURAL INFORMATION PROCESSING SYSTEMS

Maximilian Forstenhaeusler
maximilian.forstenhaeusler@bmw.de,
m.forstenhaeusler.1@research.gla.ac.uk

Daniel Külzer
daniel.kuelzer@bmwgroup.com

Christos Anagnostopoulos
Christos.Anagnostopoulos@glasgow.ac.uk



Shameem Puthiya Parambath
Sham.Puthiya@glasgow.ac.uk

Natascha Weber
natascha.weber@bmwgroup.com

Check out our work

Thank you.

University of Glasgow

Knowledge & Data Engineering Systems

BMW GROUP