T-CGAN

Conditional Generative Adversarial Network for Data Augmentation in Noisy Time Series with Irregular Sampling

20 Nov 2018, arXiv

Introduction

- For many time series data, there are **only small labeled datasets are available**.
- Solution: perform data augmentation to create synthetic data to increase the size of datasets.
- Data augmentation for time series has been limited to mainly two relatively simple techniques: **time slicing** and **time warping**.
- Time slicing: Cropping slices from time series and performing classification at the slice level.

  → Cutting the time series tends to **remove temporal correlation** in the data.
- Time warping: Warping a randomly selected slice of a time series by stretching it.

  → Not suitable for datasets whose time scale has special meaning.

Introduction

- TCGAN:
  - **Generating new irregularly-sampled time series**
  - Conditioning the generator and discriminator with the **timestamps**.
  - Assume that the time series is noisy.

Introduction

- **Experiment:**
  - Synthetic scenario: **Compare the performance** of a classifier trained with data generated by T-CGAN against the performance of the same classifier trained on the original data.
  - Real-world: consider an **unbalanced-class classification problem** and we use the T-CGAN to generate time series in the class which features the smaller training set, so as to move to a perfectly balanced setting

Technical Background

- GAN: **Generator** (capture data distribution) + **Discriminator** (identify the source of a sample)

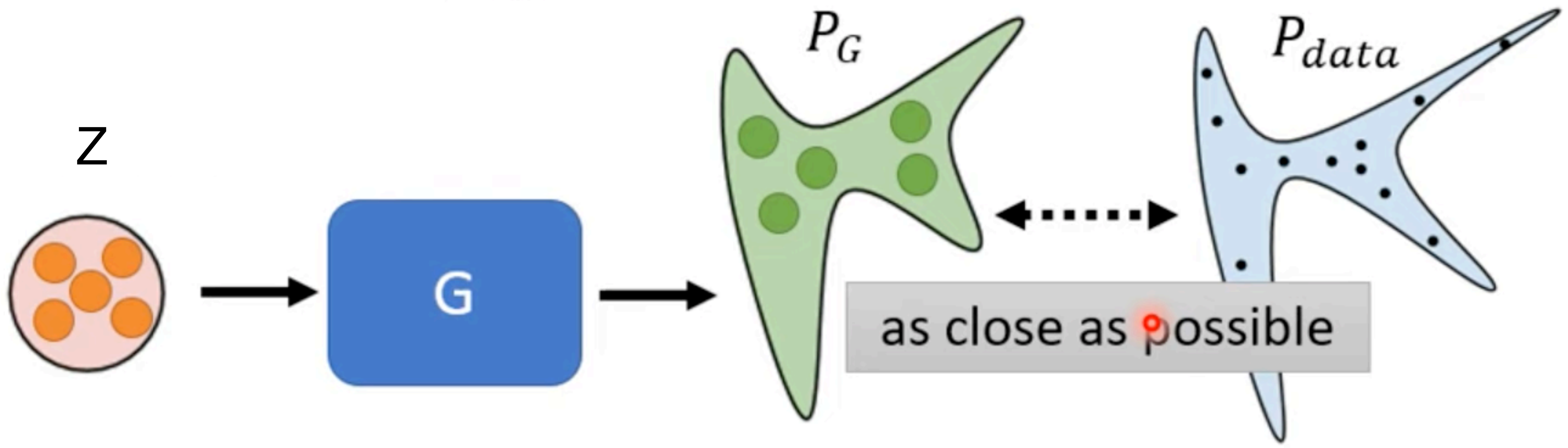
- Objective function: $G* = min_G \ max_D V(D, G)$

$$V(D, G) = E_{x \sim p_{data}}\left[logD(x)\right] + E_{x \sim p_G}\left[log(1 - D(G(x))\right]$$

- CGAN: GAN conditioned on extra information y

- Objective function: $G* = min_G \ max_D V(D, G)$

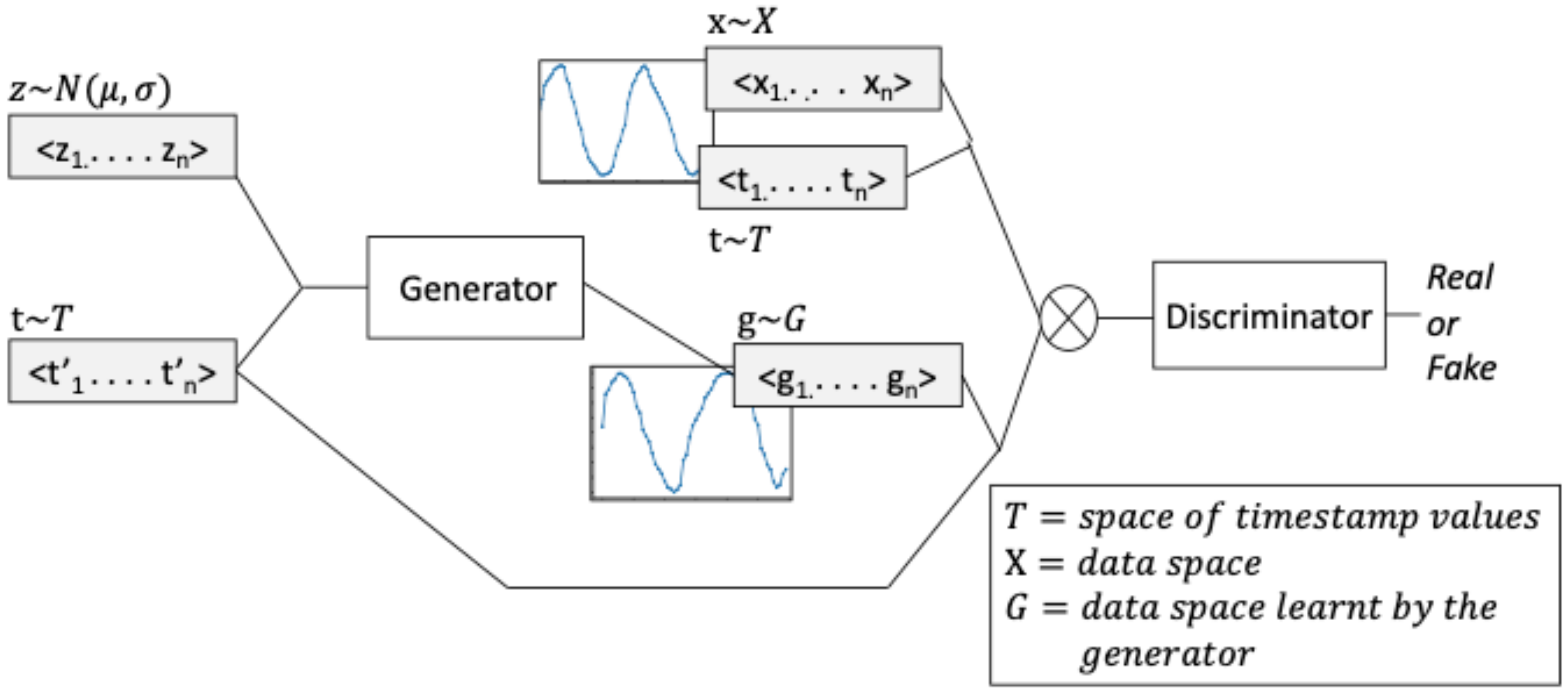
$$V(D, G) = E_{x \sim p_{data}}\left[logD(x \mid y)\right] + E_{x \sim p_G}\left[log(1 - D(G(x \mid y))\right]$$



T-CGAN Model

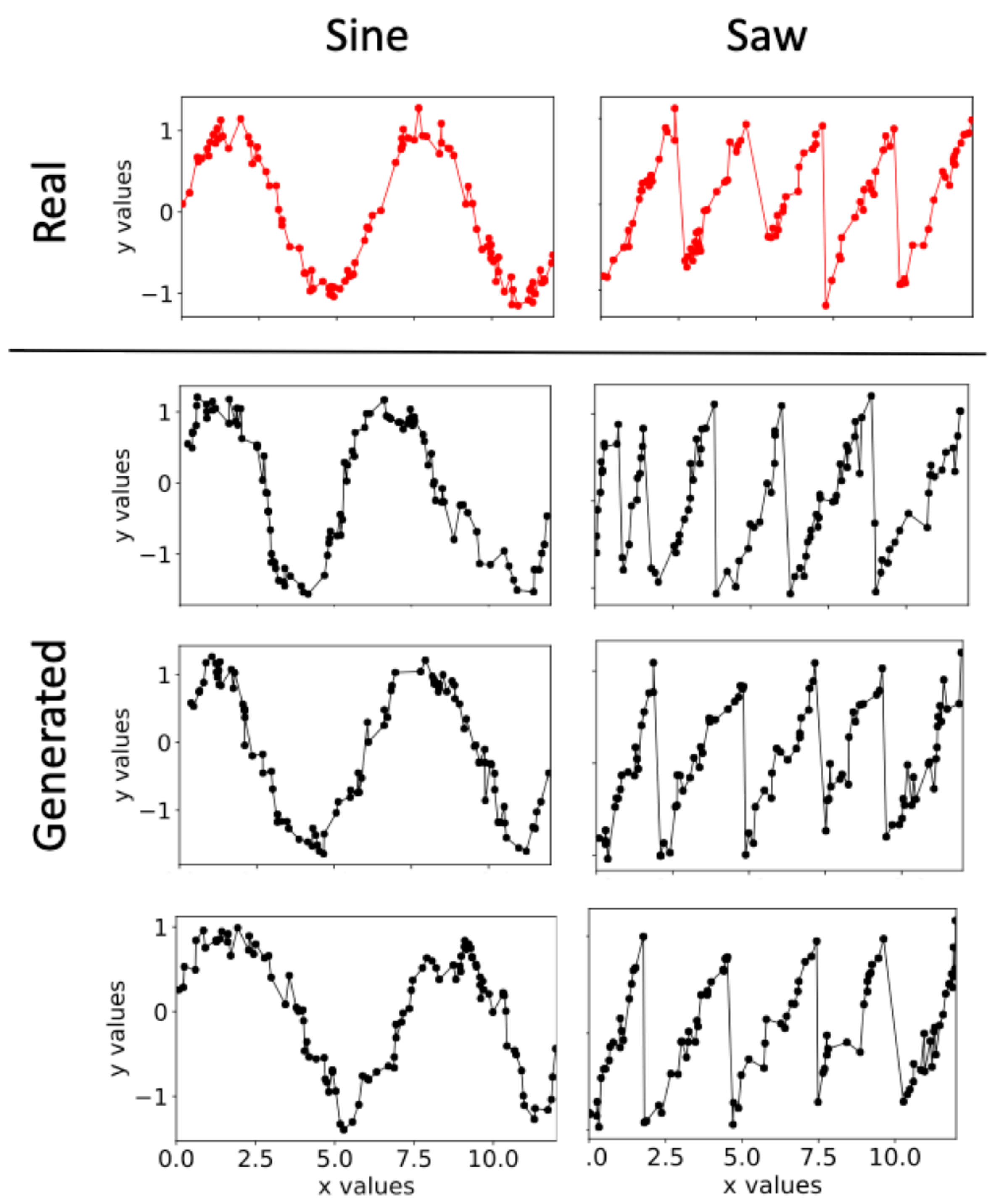
- Generator, Discriminator: two CNNs
- *Z*: a noisy space used to seed the generative model
- The objective function of T-CGAN:

$$min_G \ max_D V(D, G) = E_{x \sim p_{data}(x)} \left[log D(x \mid t) \right] + E_{z \sim p_z(z)} \left[log(1 - D(G(z \mid t))) \right]$$

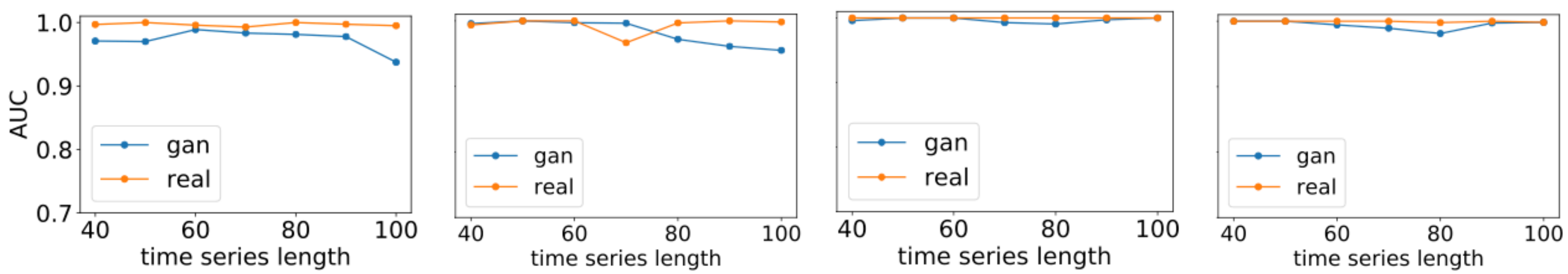


Experiments

- 10-fold randomization
- Use Area Under Receiver Operating Characteristic Curve (AUROC) to evaluate performances
- Synthetic data: sine waves and sawtooth waves



Classifier performance train on synthetic data and original data



(a) Training set size = 40 (b) Training set size = 60 (c) Training set size = 80 (d) Training set size = 100

Experiments

- Real-world data
  - Classification on regularly sampled time series.
  - Starlight curves: classify objects by their astronomical light curve.
  - Power Demand: distinguish days from summer and winter by power demand time series.
  - ECG200: distinguish heartbeat from normal and myocardial infarction by electrical activity records.

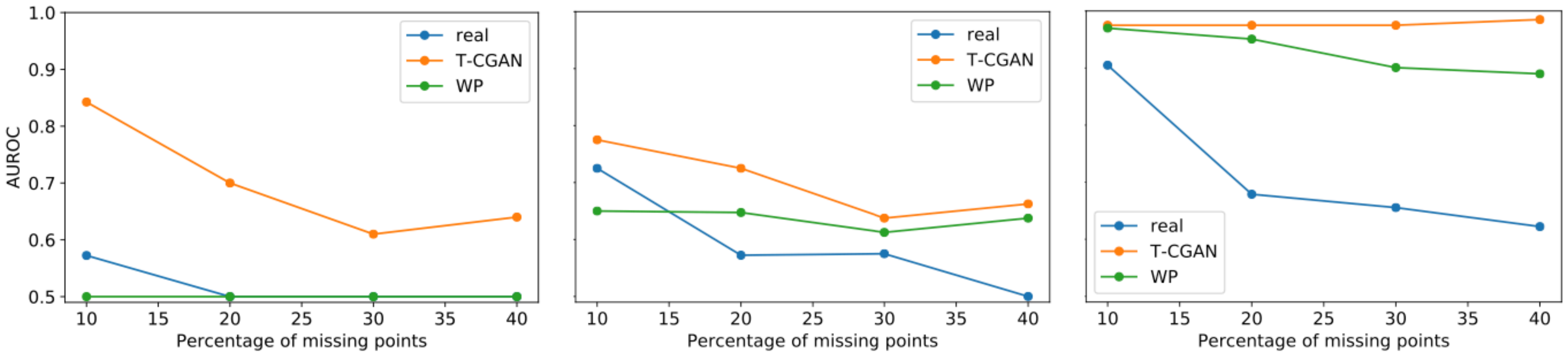
Dataset	Real data	Time Slicing	Time Warping	T-CGAN
Starlight curves	0.7127 ± 0.1371	0.7534 ± 0.0082	0.9840 ± 0.0099	$\mathbf{0.9851} \pm 0.0156$
Power Demand	0.6211 ± 0.1762	0.7152 ± 0.0932	0.7988 ± 0.0836	$\mathbf{0.8336} \pm 0.1553$
ECG200	0.7014 ± 0.0335	0.6666 ± 0.0836	0.7227 ± 0.0391	$\mathbf{0.7882} \pm 0.0122$

Experiments

- Randomly removing a certain amount of data from each series.
  - Classification on irregularly sampled time series.

Table 5: AUROC reached by each method over the different experimental scenarios, in case of irregular sampling (20% missing data, randomly selected), averaged over 10 repetitions.

Dataset	Real data	Time Slicing	Time Warping	T-CGAN
Starlight Curves	0.6798 ± 0.0222	0.5200 ± 0.0041	0.9508 ± 0.0041	$\mathbf{0.9750} \pm 0.0040$
Power Demand	0.5011 ± 0.0042	0.5020 ± 0.1240	0.5322 ± 0.0053	$\mathbf{0.6999} \pm 0.0356$
ECG200	0.5724 ± 0.2410	0.5233 ± 0.0210	0.6474 ± 0.0341	$\mathbf{0.7202} \pm 0.0546$



(a) Power Demand (b) ECG200 (c) Starlight curves

Figure 4: AUROC with varying percentage (10%, 20%, 30%, 40%) of missing values for the three datasets without augmentation (real) and with augmentation through time warping (WP) and T-CGAN (gan).