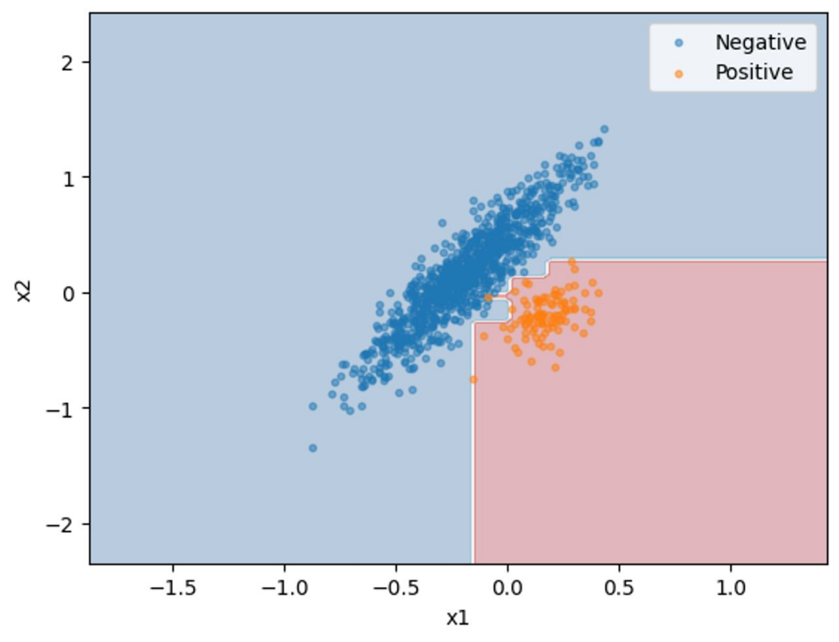
ADVANCING FAULT DETECTION IN INDUSTRIAL SYSTEMS WITH CYCLEGAN-BASED TIME SERIES DATA TRANSFORMATION

Author: Jui Chien Lin, Yu-Kai “Steven” Wang, Jonathan Li

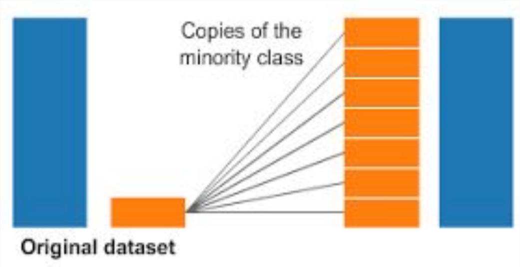
Introduction

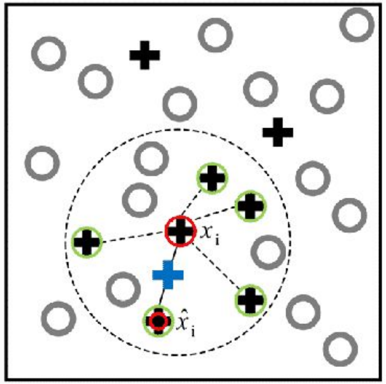
- Balanced datasets are critical for training robust and reliable machine learning models
- **Data Imbalance** is a long standing challenge for fault diagnosis of time series data from complex industrial systems
  - Normal Operational Data: inherently abundant, easy to collect
  - Faulty Operational Data: infrequent, unpredictable, hard to obtain
- Impact
  - Improper diagnosis of failure reasons
  - Overfitting on normal operational data



Background: Traditional Data Augmentation

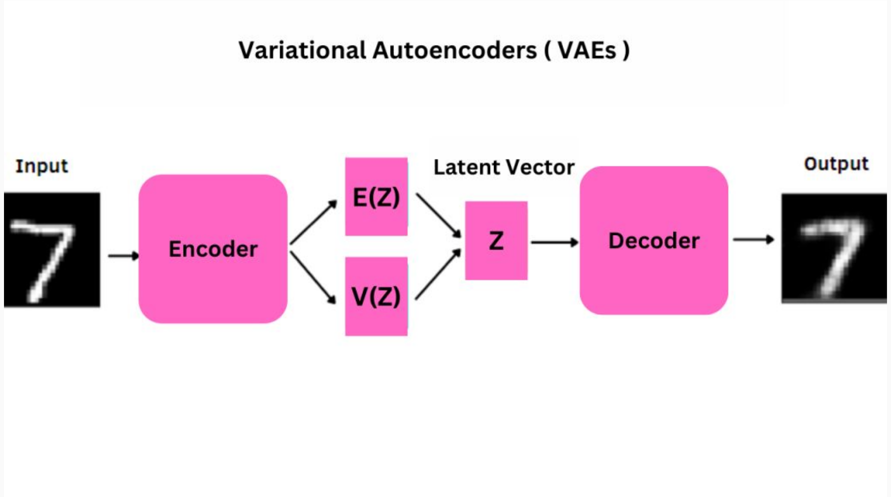
- Upsampling
  - Duplicate data samples from the minority class
  - Doesn't improve data diversity
- SMOTE
  - Data augmentation technique to rebalance minority class via linear interpolation
  - Doesn't work well with high dimensional (modern) datasets





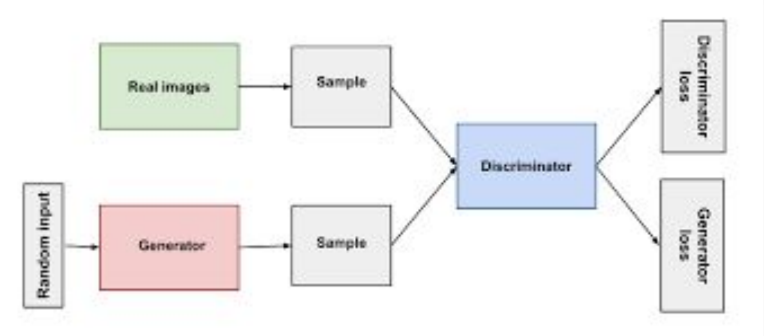
Background: VAEs

- One of the first **neural network** methods to gain traction in data augmentation community
- Encode input data into a latent probability distribution which could be sampled to get new data points
- Generated data is noisy



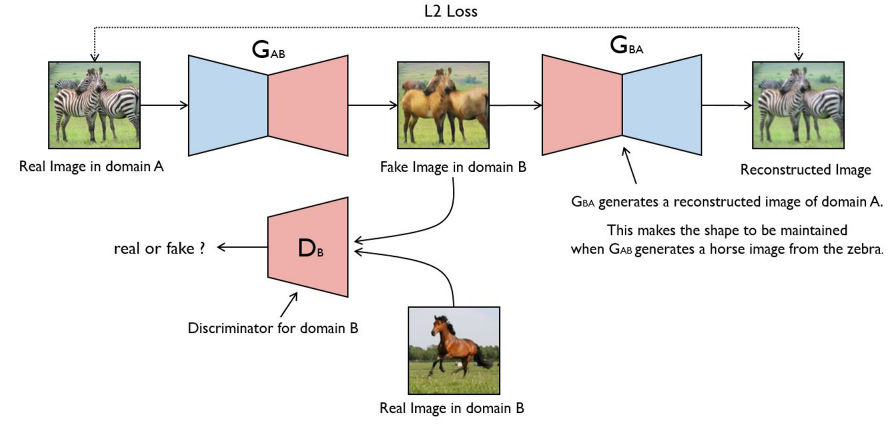
Background: GANs

- General method for high quality data synthesis that match a target distribution
- Highly used in Computer Vision community for its high quality image generation capabilities



Background: CycleGAN

- An extension of GAN originally designed to bidirectional “reversable” domain translation
- Cycle Consistency Loss
  - Minimizes reconstruction error after translating from domain to domain
- CycleGAN is a dynamic framework that can be extended to any domain



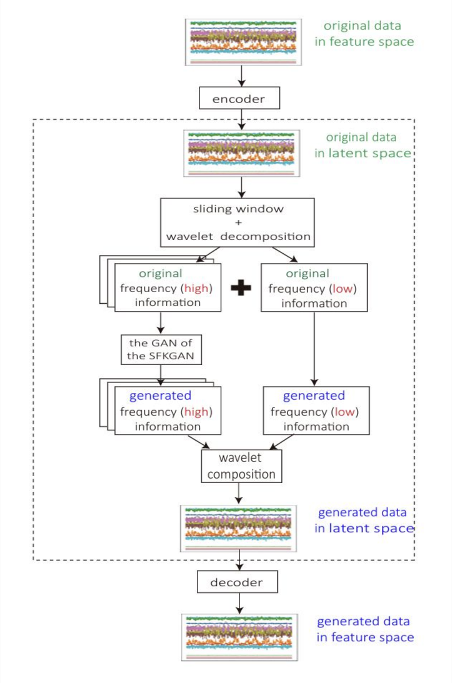
Tennessee Eastman Process Dataset (TEP)

- Process monitoring and fault detection
- Industrial time series
- For each data sample, we have 52 time series variables
- 1 normal data and 21 distinct fault scenarios, labeled as fault ID 01 ~ 21

Background: Data Augmentation for Industrial Multivariate Time Series via a Spatial and Frequency Domain Knowledge GAN (SFKGAN)

Jui Chien Lin; Fan Yang, Published in: 2022 IEEE International Symposium on Advanced Control of Industrial Processes (AdCONIP)

- Autoencoder: Compresses multivariate time series into a latent space, reducing the correlation among variables
- Wavelet Decomposition: time to frequency domain
- GAN:
  - WGAN: robust training.
  - GAN (Fully connected layers) generates synthetic high-frequency components
- Only use high frequency information of time series



Cycle Time Series GAN (CycleTSGAN)

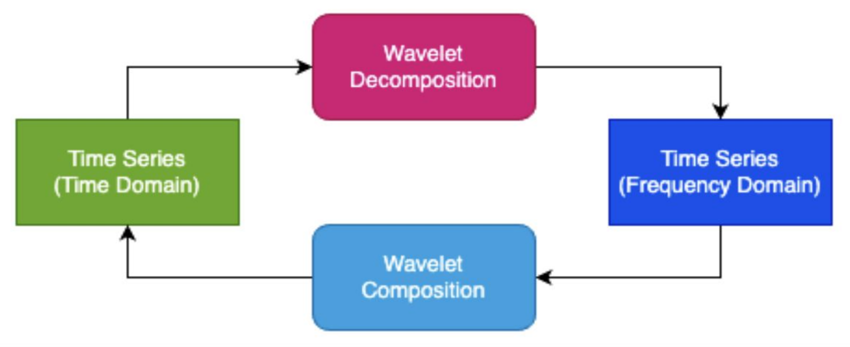
- Transformer-based Generator and Discriminator
- Wasserstein loss for both generator and discriminator
- L1 cycle consistency loss
- highlight:
    - Unlike SFKGAN, which overlooks the inherent properties of abundant normal data, CycleTSGAN synthesizes faulty data by transforming normal data rather than mapping Gaussian noise to the target distribution.
    - CycleTSGAN is based on the premise that normal and faulty data share common characteristics, as they originate from the same system operating under different conditions.

Preprocessing: Wavelet Decomposition

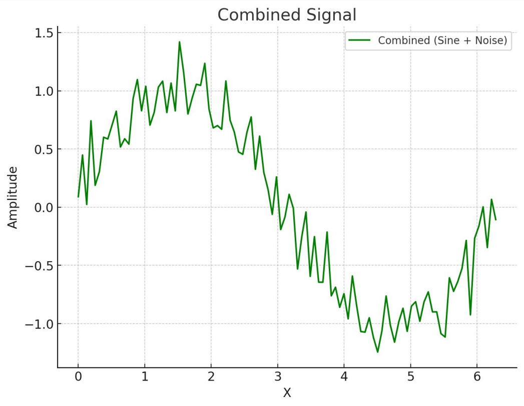
- Wavelet decomposition
  - High frequency components
  - Low frequency components
- Multi-level representation of the signal in frequency domain

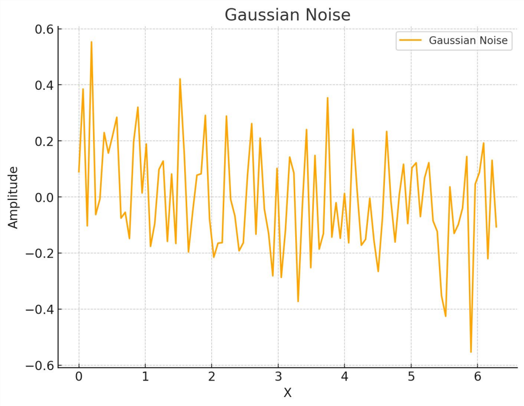
$$y_{\text{high}}[n] = \sum_{k=-\infty}^{+\infty} x[k] \cdot g[2k-n]$$

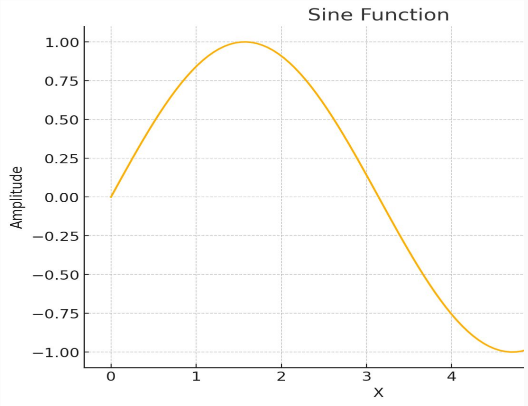
$$y_{\text{low}}[n] = \sum_{k=-\infty}^{+\infty} x[k] \cdot h[2k-n]$$



Wavelet Decomposition Visualization







Original Time series

High frequency info

low frequency info

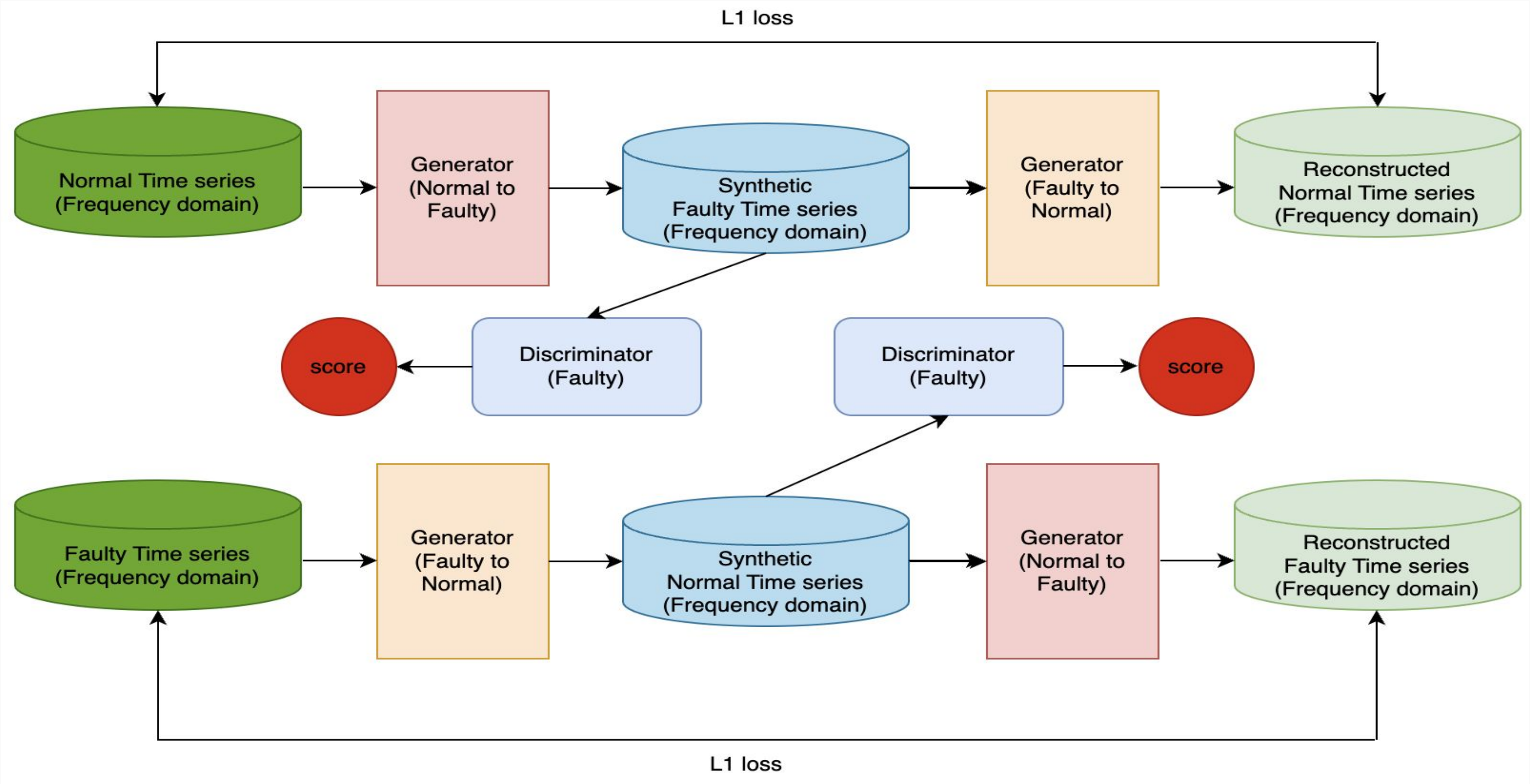
Wasserstein Loss (Earth Movers Distance)

- Improves on vanilla GANs by addressing instability in measuring distances between disjoint data distributions
- Wasserstein Distance Formula:

$$W(P_1, P_2) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{\mathbf{x} \sim P_1}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim P_2}[f(\mathbf{y})]$$

- Wasserstein Loss:

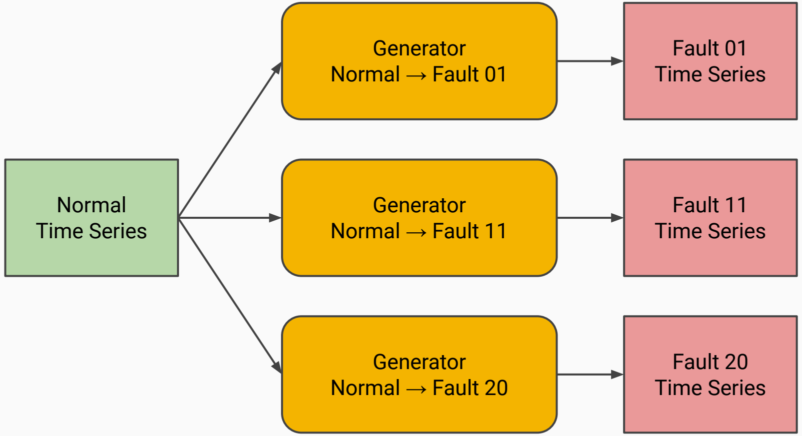
$$\min_D \max_G \mathbb{E}_{\tilde{\mathbf{x}} \sim P_{\text{real}}}[D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{z} \sim P_{\mathbf{z}}}[D(G(z)] + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim P_{\hat{\text{real}}}}(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2$$

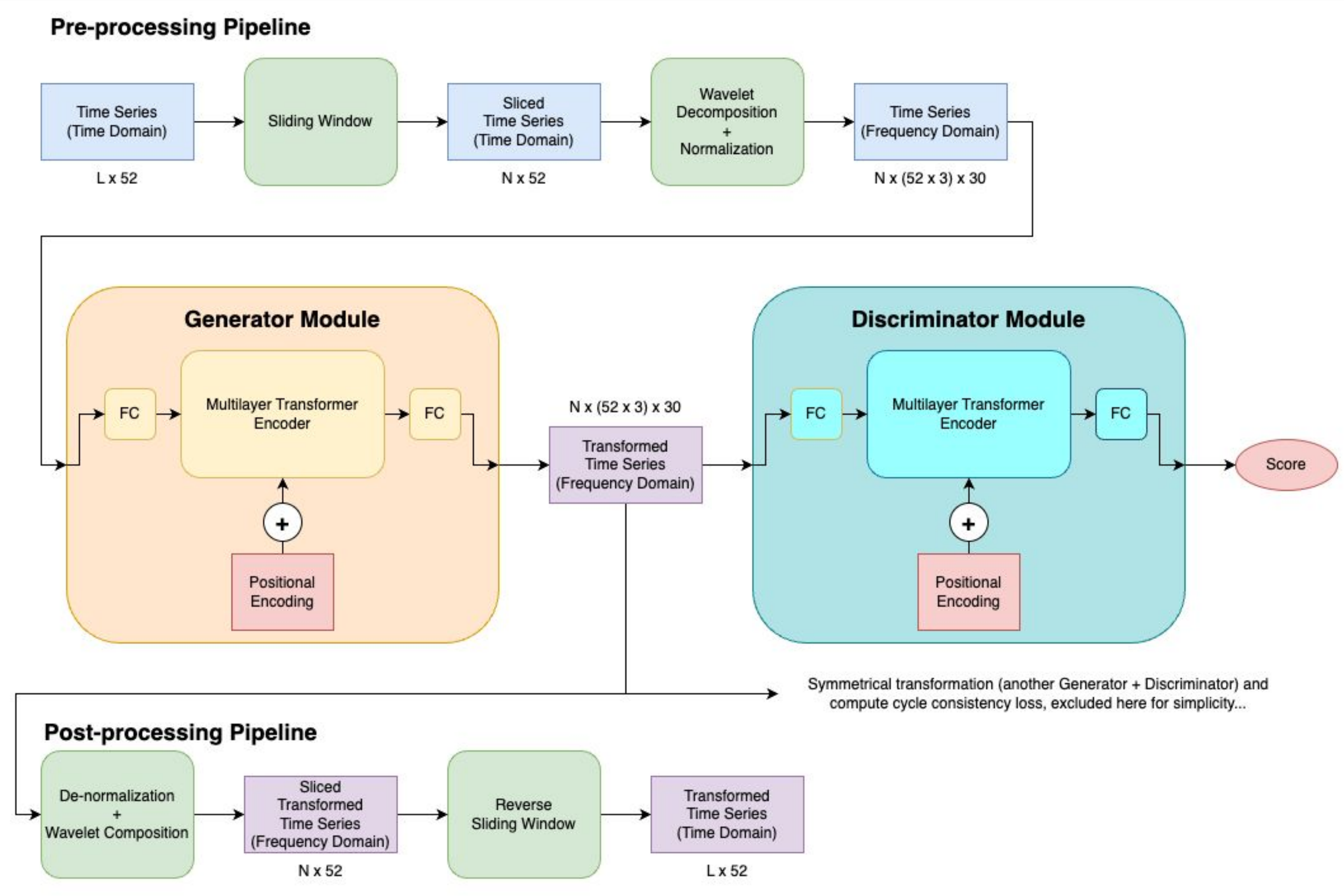


CycleTSGAN Model Architecture

Training CycleTSGAN

- Trained three models to transform normal data to fault ID 01, 11, and 20
- 20k epochs using AdamW optimizer
- 1k warmup epoch, 100k total epoch
- learning rate 0.0001, with a constant then linear decay scheduler
- L1 cycle consistency + L1 transform loss + Wasserstein adversarial loss





Details of the Training Process

Model Evaluation

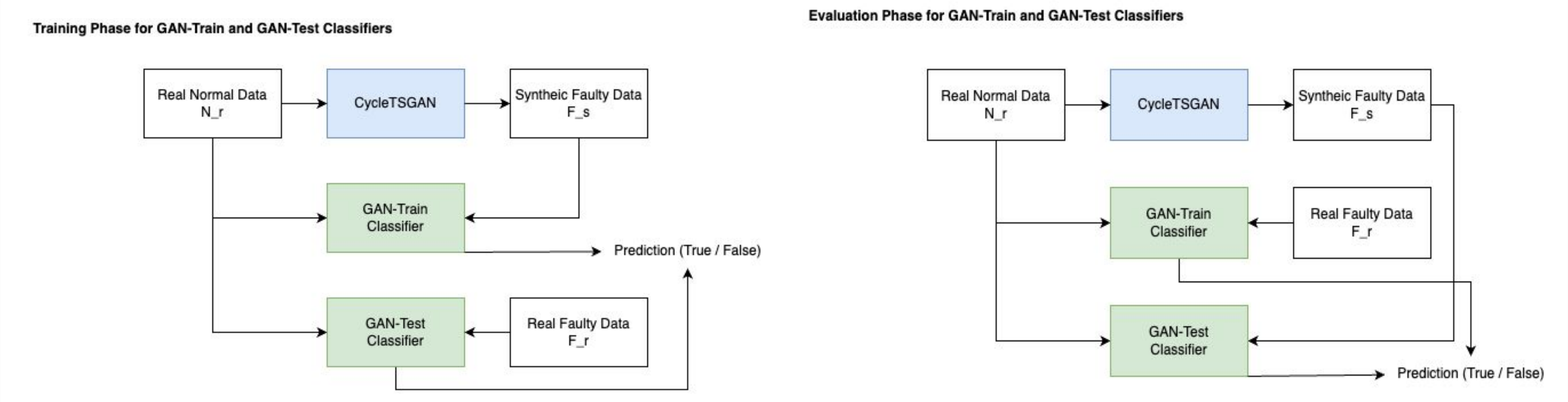
GAN-Train & GAN-Test

- Measure how well CycleTSGAN captures the target distribution
- Train two binary classifiers: GAN-Train & GAN-Test
  - GAN-Train: Train on synthetic data, eval on real data
  - GAN-Test: Train on real data, eval on synthetic data

How good is my GAN?

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari

Inria*



GAN-Train measures “recall”, GAN-Test measures “precision”

GAN-Train & GAN-Test Workflow

GAN-Train & GAN-Test Results

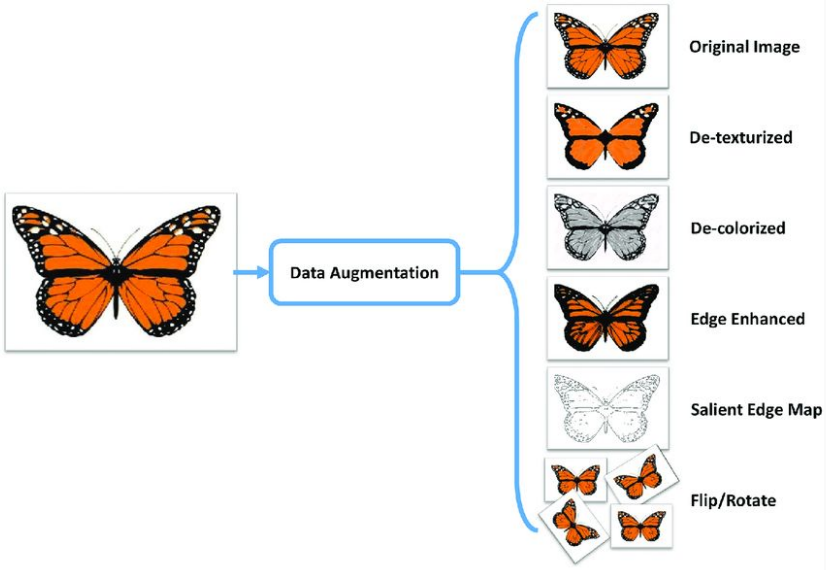
- All GAN-Train & GAN-Test trained until perfect accuracy
- Possible mode collapse on fault id 01 (high precision low recall)
- Slight mode collapse on fault ID 11 (high precision, ok recall)
- Good results for fault ID 20 (high precision, high recall)

Fault ID	GAN-Train Training Accuracy	GAN-Train Evaluation Accuracy
01	1.000	0.500
11	1.000	0.779
20	1.000	1.000

Fault ID	GAN-Test Training Accuracy	GAN-Test Evaluation Accuracy
01	1.000	1.000
11	1.000	1.000
20	1.000	1.000

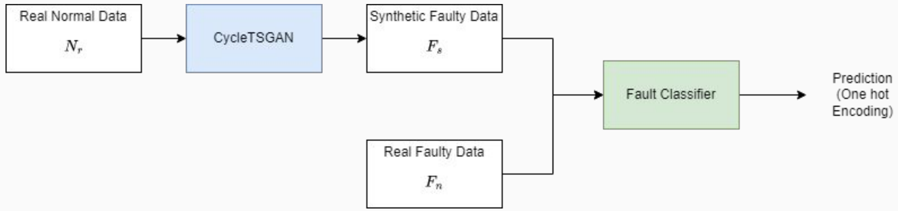
Data Augmentation

- Generated data can be used to mediate data imbalance issues
- If CycleTSGAN generates new data from the faulty operational data distribution, we should see an improvement on a classifiers test accuracy



Data Augmentation Experiment Details

- Imbalance dataset
  - 300 Real Fault ID 01, 300 Real Fault ID 11, and 50 Real Fault ID 20 samples
- Data Augmentation Rebalancing
  - A new dataset will be created by adding 50 synthetic Fault ID 20 samples
  - Synthetic data will be iteratively added at increments of 50 until Fault ID 20 has 300 total samples
- 10% of training data is used for validation
- Dataset with 100 samples from each fault id will be reserved for testing
- A LSTM + MLP classifier will be trained on each training dataset and evaluated on the test set



Data Augmentation Results

- F1 Score for Fault 20 and Overall Accuracy increase when more synthetic Fault ID 20 samples are added
- The best model trained on each dataset had a 100% training/validation accuracy
- Data augmentation is important not only for rebalancing the dataset, but to gauge the performance of a model on unseen data

Synthetic Fault ID 20 Samples	F1-Scores			Overall Accuracy
	Fault 01	Fault 11	Fault 20	
0	1.00	0.85	**0.79**	0.88
50	1.00	0.88	**0.85**	0.91
100	1.00	0.93	**0.92**	0.95
150	1.00	0.95	**0.95**	0.97
200	1.00	0.93	**0.93**	0.95
250	1.00	0.97	**0.96**	0.98

Conclusion

- CycleTSGAN provides an innovative solution for data augmentation in time series data by leveraging CycleGAN based transformations and wavelet decomposition to generate high quality synthetic industrial time series data
- Our approach utilizes the abundance of normal operational data to mitigate the lack of faulty operational data
- Results on TEP Dataset highlights the effectiveness of CycleTSGAN
  - Clear gains in classification accuracy and robustness through adding synthetic data to minority classes of unbalanced datasets
  - GAN-Train and GAN-Test provide comprehensive insights into the quality and utility of the synthetic data
- Future Work
  - Mitigating mode collapse for certain fault types
  - Scaling framework to more complex datasets
  - Incorporate adaptive mechanisms to further enhance realism and diversity of generated data

Questions