Vision Model as Foundational Timeseries forecaster

Korede Bishi,Amily Chowdhury, Farzaneh Sadati, Noyon Dey
School of Computing, The University of Georgia
{korede.bishi,amily.chowdhury, farzaneh.saadati, noyon.dey}@uga.edu

Abstract

The use of foundation models for time series forecasting (TSF) has become a popular choice. However, existing trends primarily rely on developing large-scale datasets to train TSF models. Recent advances have shifted towards leveraging natural image reconstruction processes for TSF, showing improved performance over traditional methods. Notably, prior work has mostly focused on univariate time series data using this approach due to the interdependency of sub-features with time in multivariate time series. In this work, we address this gap by developing a model for multivariate time series forecasting, by following ideas from computer vision. In this work, we propose an image reconstruction-based technique where we treated the leftmost 80% of the image as the historical data and reconstruct the rightmost 20% of the image, aligning with the left-to-right TSF problem. We followed a latent representation learning technique using a variational autoencoder (VAE). We transformed multivariate time series data into an image grid where each grid corresponds to a variable from the multivariate data. These grid-based images were then used to train VAE to learn the latent representation. Experimental results are impressive compared to the masked autoencoder (MAE)-based approach.

1. Introduction

Time series forecasting (TSF) refers to predicting future values based on historical data and is a critical task in fields such as finance, healthcare, and climate modeling, where predicting future values based on historical data is essential. TSF has implications in various fields such as social, economic, and so on. For this reason, there has been a constant endeavor to improve TSF tasks ranging from univariate to multivariate time series signals.

Traditional methods such as ARIMA, ETS, and state-space models were the initial approaches to TSFs but with time, there is more data complexity and non-linearity. Traditional models often fail when faced with these challenges. To mitigate these challenges, more improved deep learning (DL) models are proposed such as recurrent neural networks (RNN) [11], long short-term memory (LSTM) [7], transformers [14], and so on.

These traditional and comparatively newer DL-based approaches are mostly text or token-based approaches that consider TSF as a sequential text or data processing. These traditional approaches to TSF involve deep learning models like LSTMs or Transformers, which perform well on large-scale datasets but can struggle with multivariate data due to their complexity and resource requirements [7], [14]. Hence, there is a need to recognize and solve these complexities with more efficient approaches for TSF, and this is seen in the work of Chen et. al [3]. In this work, the authors tried an out-of-the-box approach named *VisionTS* to solve univariate TSF tasks using a computer vision approach and found an improved result over the state-of-the-art (SOTA) TSF task.

Recent advancements in computer vision, particularly the use of Masked Autoencoders (MAE) [6] for images and video frame reconstruction, offer an exciting new avenue for improving TSF and this has been successfully investigated in [3]. These models have shown significant promise in capturing spatial and temporal relationships but have primarily been applied to uni-variate data. Applying MAE, typically used for videos, to multivariate time series forecasting is a novel approach. In this project, we aim to bridge this gap by applying VideoMAE to multivariate time series forecasting.

Though there is already research work on univariate TSF using MAE [3], to the best of our knowledge, there is no work on multivariate TSF with a computer vision approach. There are complexities such as high dimensionality, high variability, data sparsity and missing values, irregular sampling, noise, and so on, that significantly challenge TSF with the computer vision approach.

Hence, in this work, we working to bridge this gap and we are doing TSF with multivariate time series data using a computer vision approach. To be specific, we are following an image reconstruction-based approach to predict the rightmost 20% of an image, thereby aligning with the TSF left-to-right prediction task. We record each variable from

the multivariate time series data as a line graph plot following ViTST [9]. We follow the same procedure to record each variable as a line graph plot and eventually assemble all the variables' plots into a grid-based, forming an image. This process eases the training procedure for the multivariate time series data with a computer vision approach. We aimed to learn the latent representation from these training images with the help of a variational autoencoder (VAE) [8] so that we can predict or reconstruct any images that has been transformed into a line graph-based image.

By converting multivariate signals into image frames, predicting using VAE, and reconstructing the time series, we hope to provide a more robust forecasting model. Compared to traditional TSF models, this approach is expected to provide more accurate forecasts and greater generalization across domains. The remainder of this report is organized as follows: section 2 provides related work, section 3 describes used datasets, the proposed methodology is presented in section 4, evaluation results are in section 5, and eventually, this work is concluded in section 6

2. Related work

Traditional TSF methods like LSTMs and Transformers perform well on univariate data but struggle with complex multivariate datasets [7]. MAE was developed for image reconstruction [6], and later adapted for video data with VideoMAE, capturing both spatial and temporal information [13]. VideoMAE [12] has shown success in computer vision through self-supervised learning [13]. These computer vision approaches have been explored in other fields such as TSF. Such is seen in VisionTS [3] where MAE is used to TSF task using univariate time series data. However, its application to multivariate time series forecasting remains largely unexplored.

Applications of computer vision in TSF, especially in multivariate TSF in its earlier phase. Recent works such as ViTST [9] have shown the use of transformer-based [14] image classification model (vision transformer) [4] for multivariate time series classification. Though ViTST is an interesting and earlier approach to applying an image transformer to a time series, application in multivariate TSF is yet to done. Hense, in this work, we aim to work in this novel approach where we plan to forecast multivariate time series data using a computer vision approach. To be specific, we aim to forecast or predict multivariate time series data using an image reconstruction-based approach using VAE.

3. Used Datasets

This study uses two datasets for training and one for testing. For training, we used the P19 [10] and P12 [5] human activity sensor datasets. The P12 dataset has about 12,000 images after transformation. On the other hand, the P19 dataset produces about 39,000 images. For testing, we used a one-month vehicle trajectory dataset downloaded from PeMS [1], a traffic monitoring platform. After cleaning and preparation, as described in section 4, the time series data was turned into line-graph images, resulting in about 1,200 test samples.

4. Proposed Methodology

Conventional time series forecasting methodologies often rely on temporal models that infer future values from a fixed historical window. While effective, these approaches do not fully exploit the spatially coherent representations that emerge when complex multivariate signals are mapped into image formats. Inspired by recent advances in image-based modeling for time series tasks [3], we propose treating multivariate time series data as structured image-like inputs, thereby leveraging established image modeling techniques to learn latent representations conducive to forecasting. Figure 1 demonstrates our proposed approach where we transform multivariate time series data into an image format, learn its latent representation, reconstruct it, and eventually reconstruct the numerical data from the reconstructed image. The following subsections describes our methodology in details.

4.1. From Raw Data to Image Representations

In order to convert our raw multivariate data into an image format, we followed one recent work in multivariate time series data that adapted an image-based technique [9] for multivariate time series classification. Following the guidelines, we developed and tested our methodology on P12 and P19 datasets as described in section 3.

4.1.1 Training Datasets: P12 and P19

For training, we employed the P12 and P19 human activity sensor datasets as described in [9]. These datasets contain a larger number of variables, each represented as a time series line graph. To form a single image from these multivariate time series, line graphs for all variables are arranged on a grid. Following the approach in [9], a square or near square layout is chosen to accommodate the total number of variables. For example, P19 and P12 datasets each with 34 and 36 variables, respectively use a 6×6 grid, where empty grid cells remain blank if fewer than 36 variables are present (Figure 2a). The order of variables within the grid can be sorted by criteria such as missing ratios or other domain-informed heuristics. This image creation strategy, as exemplified by P12 and P19, effectively transforms complex multivariate time series into structured images suitable for downstream tasks. By standardizing both the plotting and

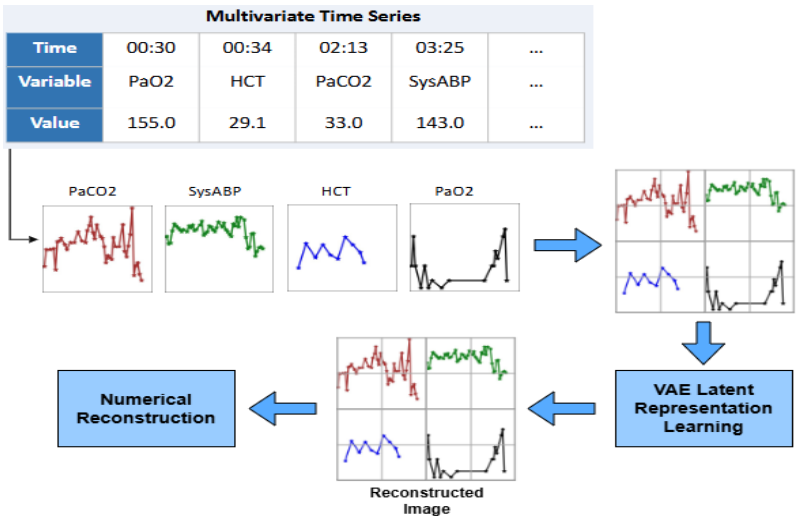
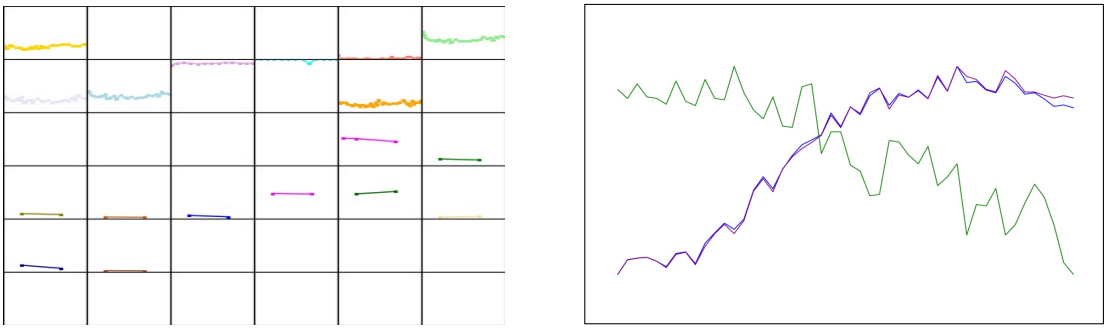


Figure 1. **Overall framework of the proposed method that predicts multivariate time series data using VAE model.**



(a) P12 dataset sample.

(b) PeMS normalized dataset sample.

Figure 2. **Used dataset samples.**

arrangement processes, this approach simplifies the ingestion of time series data by image-based models, facilitating latent representation learning and subsequent forecasting.

4.1.2 Testing Dataset: PeMS Dataset

We obtained one month of traffic data from Caltrans, focusing on a straight segment of the US101 in San Francisco equipped with 20 sensors embedded across 4 lanes [1]. This configuration provided a rich multivariate time series, capturing flow, occupancy, and speed measurements at 5-minute intervals. To improve computational efficiency and harmonize the forecasting horizon, the raw data was aggregated into 15-minute intervals.

After cleaning and retaining only the three key variables flow, occupancy, and speed, we normalized the resulting time series to ensure comparability across sensors and features. We then plotted each sensor's normalized data as line graphs over distinct day and night intervals (6:00 AM–6:00 PM and 6:00 PM-6:00 AM, respectively, as shown in Figure 2b). This segmentation allowed us to capture diurnal variations in traffic patterns. In total, we generated approximately 1,160 images, each encoding the temporal dynamics of multiple sensors within a single daily or nightly snapshot.

4.2. VAE-based Latent Representation Learning

To forecast multivariate TS data from transformed images, we followed an image reconstruction-based approach similar to VisionTS [3]. The core idea was to predict or reconstruct the rightmost 20% of the input image based on the leftmost 80% of the image. This is similar to the TSF task where the rightmost part (i.e., future information) is predicted based on the leftmost visible part (i.e., past information).

Hence, we adopted to learn the latent representation from grid-based transformed image and selected VAE [8] for this task. The idea was to learn the latent representation of the transformed image and generalize it on general multivariate data following an image reconstruction approach. To be precise, we aim to predict or reconstruct the rightmost 20%

of that image based on the leftmost 80% visible parts.

We describe our framework in Figure 1. In addition to this approach, we initially started with MAE, then VAE, and at the end, we also experimented with denoising autoencoder (DAE).

4.3. Numerical Reconstruction

Our proposed approach for reconstructing the original time-series data involves a two-step process. First, we envision reverse engineering the visualization pipeline by mapping the normalized NumPy arrays (designed for better line plot visibility) back to their original numerical values. This would involve interpreting the graphical patterns to recover the underlying data. Second, we propose validating the reconstruction by comparing the reverse-engineered data with the original time series dataset to assess accuracy. While this concept remains unimplemented due to practical challenges, it outlines a pathway for ensuring the recoverability of raw numerical data from visual representations.

5. Experimental Evaluation

We experimented with three model architectures: MAE, VAE, and DAE.

5.1. Experiments

For pixel-level reconstruction experiments with MAE, we followed three different approaches with three versions of pre-trained MAE. We fine-tuned MAE with their pre-trained weights and experimented with base, large, and huge versions of the pre-trained MAE model from Huggingface [2]. We used both P12 and P19 datasets separately to train MAE versions, however the reconstructed results were not optimally aligned with the ground truths as shown in Figure 3b. Though MAE had better mean PSNR (26.89) value than VAE (25.30), it was not perceptually closer to the ground truths. We then experimented with various parameter changes with the MAE model versions whether we have an improvement or not. In the end, we did not get a very good reconstruction; the reconstructed rightmost 20% area was mostly smoothed out with image pixels scattered which did not represent any meaningful structure when we visually compare with both the ground truth and the leftmost 80% to the rightmost 20% reconstructed part. This is probably due to the specific masking strategy and inherent training data characteristics. We changed MAE's random masking to the rightmost 20% masking in our training phase. This might not align well with the pre-trained model as it was trained with a random masking strategy. In addition, the training images had mostly white regions and signals represented as line plots. Hence, the white area dominates and the model did not learn the pixel level details from the line plot. In addition, image reconstruction based on pixel-level reconstruction is an extremely difficult task requiring larger datasets. We believe these are the reasons for which we did not get a very good result with MAE.

With the poor reconstruction results from the MAE-based approach, we then moved towards a latent representation learning approach and reconstruction based on that feature learning. The idea was to learn the latent representation of the line graphs inside the image and to generalize with similar datasets and be able to reconstruct or predict those scenarios. To this end, we adapted VAE and fine-tuned it on our training data. We got comparatively better-reconstructed image parts (mean LPIPS of 0.07 compared to MAE (0.0758)) that we were looking for as shown in Figure 3c. Lastly, we experimented with DAE but found the VAE had the best results.

5.2. Evaluation Metrics

We followed both quantitative and qualitative performance measurements of the reconstructed image. For the quantitative approach, we have used peak signal-to-noise ratio (PSNR) and mutual information (MI). We used learned perceptual image patch similarity (LPIPS) for the qualitative one.

5.2.1 PSNR

PSNR measures the difference (error) between original and reconstructed images expressed as a ratio decibels (dB). A higher PSNR value means a better quality image reconstruction. It is calculated based on the Mean Squared Error (MSE) between the original image I and the reconstructed image I':

$$\text{MSE} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(I(i,j) - I'(i,j) \right)^2$$

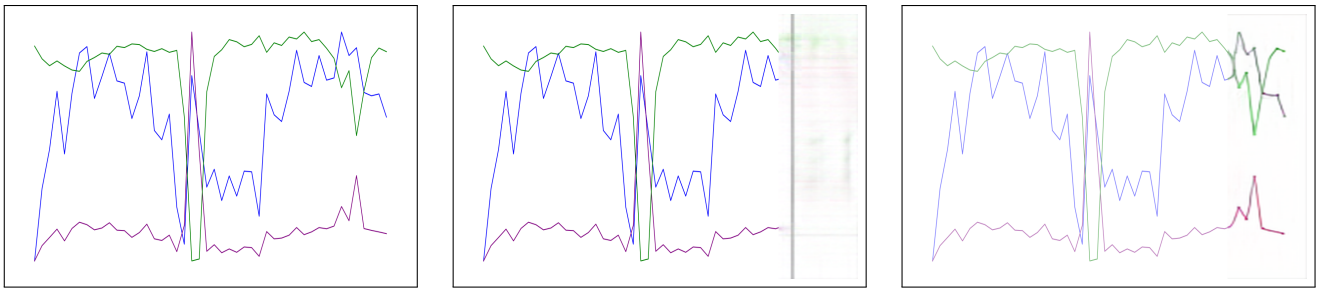
where:$I(i,j)$: Pixel intensity of the original image at position (i,j). $I'(i,j)$: Pixel intensity of the reconstructed image at position (i,j) m,n: Dimensions of the image

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right)$$

5.2.2 MI

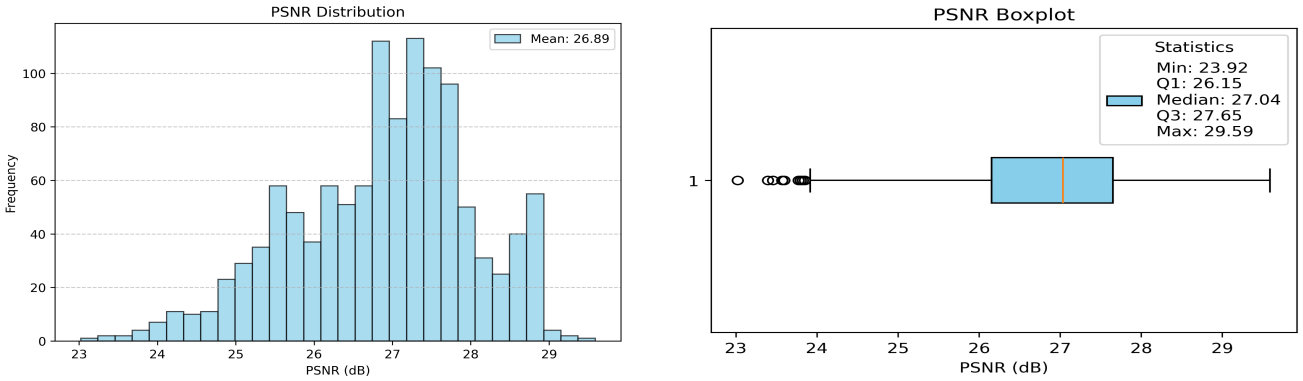
MI tells you how much information about the original image is retained in the reconstructed version of that image i.e., how information is shared between the two images. It quantifies the pixel intensities between two images and provides an evaluation of the reconstruction. A higher MI is better quality reconstruction.

For the original image I_{orig} and the reconstructed image I_{recon}, the mutual information is given by:

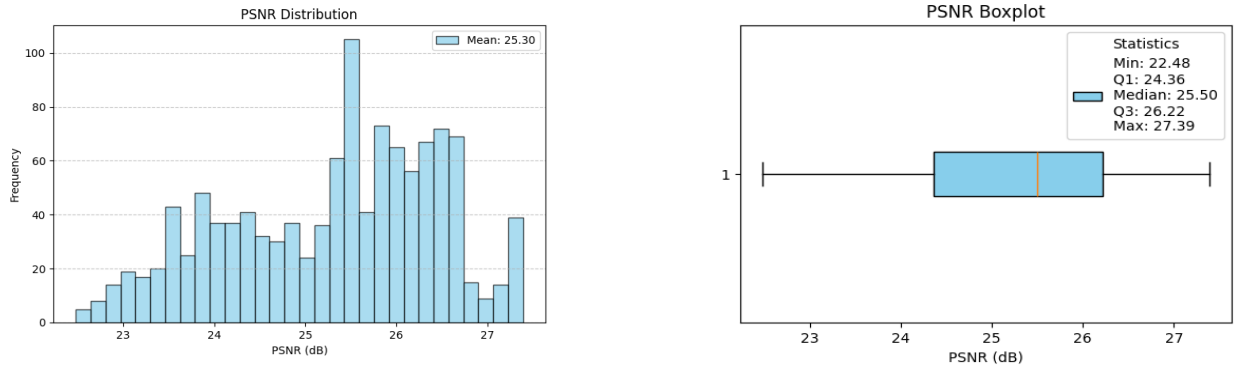


(a) Groundtruth image. (b) MAE reconstructed image. (c) VAE reconstructed image.

Figure 3. **Ground truth, VAE, and MAE generated images.**



(a) PSNR histogram distribution of MAE. (b) PSNR boxplot distribution of MAE.



(c) PSNR histogram distribution of VAE. (d) PSNR boxplot distribution of VAE.

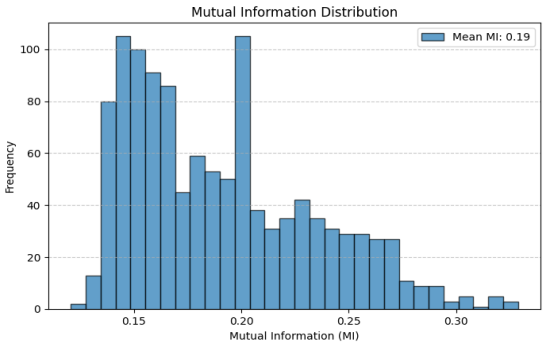
Figure 4. **Pixel level similarity performance of MAE and VAE using PSNR evaluation.**

$$I(I_{\text{orig}}; I_{\text{recon}}) = \sum_{i,j} p(i,j) \log \left(\frac{p(i,j)}{p(i)p(j)} \right)$$

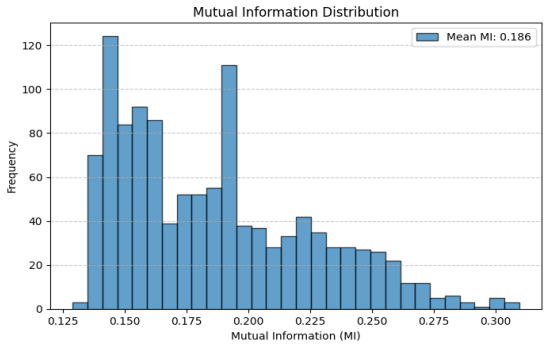
Where: $p(i,j)$: Joint probability of pixel intensity i in I_{orig} and j in I_{recon}. $p(i)$: Marginal probability of pixel intensity i in I_{orig}. $p(j)$: Marginal probability of pixel intensity j in I_{recon}.

5.2.3 LPIPS

LPIPS is a qualitative evaluation metric that captures the learned perceptual similarity between original and reconstructed images in a way that aligns with human perception [15]. It uses a deep neural network to capture the perceptual differences between images that align with human perception of visual similarity. A lower LPIPS value means better perceptual similarity.

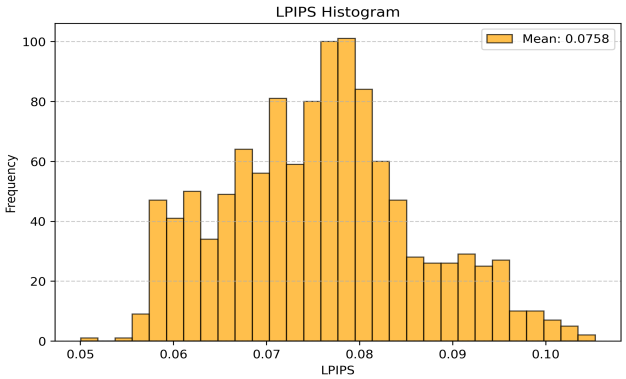


(a) MI histogram distribution of MAE.

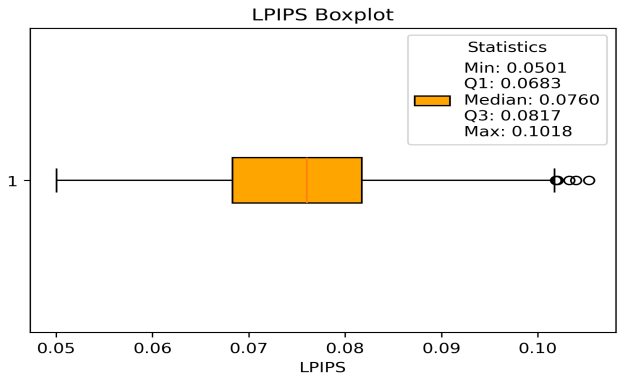


(b) MI histogram distribution of VAE.

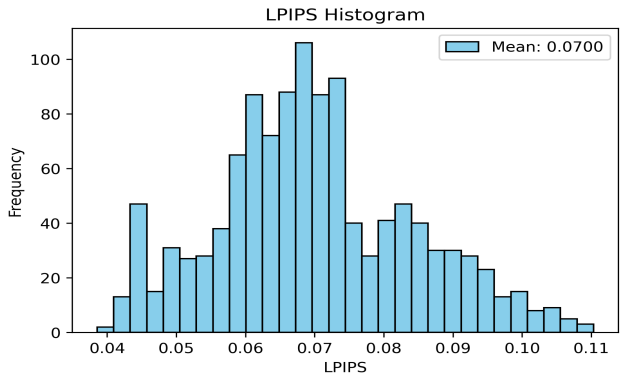
Figure 5. **Shared information performance of MAE and VAE using MI evaluation.**



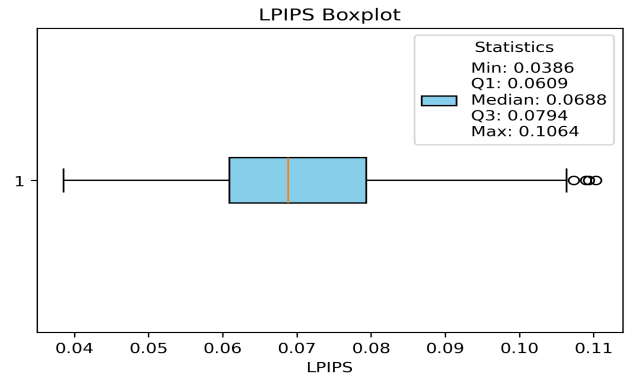
(a) LPIPS histogram distribution of MAE.



(b) LPIPS boxplot distribution of MAE.



(c) LPIPS histogram distribution of VAE.



(d) LPIPS boxplot distribution of VAE.

Figure 6. **Visual similarity performance of MAE and VAE using LPIPS evaluation.**

5.3. Quantitative Performance

We evaluated on the test dataset (1160 images) with PSNR and MI.

Based on Figures 4a and 4c, it is evident that MAE has a better mean PSNR value (i.e., 26.89) than the VAE (i.e., 25.30). Though the MAE-based reconstructions are not visually pleasant, they have pixels scattered in various places in the reconstructed part that correspond to the ground truth pixel intensities (Figure 3b). This is why we got a higher mean PSNR value for MAE-based reconstruction. In addition, from the boxplot of MAE (Figure 4b), we can also see that it has some outliers and a portion of the images

has significantly lower PSNR than the mean. On the contrary, the VAE PSNR boxplot (Figure 4d) does not show these divergences which also shows that the majority of the reconstructions are aligning with the mean.

We also evaluated another quantitative performance using the MI evaluation metric where a higher MI value is expected to have a better reconstruction. Figures 5a and 5b refer to MAE and VAE approaches, respectively. Here, again we can see a similar trend of higher MI value in the MAE-based approach (mean = 0.1908) than the VAE-based approach (mean = 0.18664). This is due to the pixel intensity-based approach in MI and also pixel intensity spread in the MAE-based reconstruction as discussed earlier. For this reason, we got higher values both in PSNR and MI for the MAE-based reconstruction than the VAE-based approach despite having lower visual and structural similarity. For this reason, we adopted a perceptual similarity measurement approach in the qualitative performance measurement.

5.4. Qualitative Perceptual Performance

For the perceptual similarity measurement, we have used LPIPS where lower values of LPIPS mean better reconstruction perceptually. Figures 6a and 6c show the histogram distribution of LPIPS values of the test dataset with frequencies. It is evident that the mean LPIPS is lower in the VAE-based approach (mean = 0.07) than in the MAE-based approach (mean = 0.0758). This also means that VAE has better reconstruction than the MAE-based approach and this is also true if we look at the reconstructed image's structure and quality.

5.5. Discussion

Preparing input data for image reconstruction from the multivariate time series data is a challenging task. Initially, we planned to use a video-based reconstruction approach using VideoMAE [12] where we planned to treat each variable as a video frame generated following the VisionTS [3] framework. Following the VisionTS framework for each variable to generate a frame is a costly process both in terms of time and space. In addition, numerical reconstruction from video frames was also extremely difficult. Hence, we moved to a different approach where we worked with image-based reconstruction with multivariate data following the work of ViTST [9]. This approach simplified our input processing and reconstruction procedure. However, we still faced issues in numerical reconstruction from the reconstructed image.

For image reconstruction, we have taken VAE over MAE and DAE (i.e., latent representation or feature learning over pixel-based reconstruction). Based on the experiments and their results, we found an improved perceptual and structural reconstruction from VAE. However, there is still a scope for improvement as we did not get a high mean PSNR and LPIPS values for VAE. We plan to improve this situation with more diverse and larger multivariate time series transformed data in the future with a more complex model.

We experienced and tackled various challenges from processing multivariate time series data into images, and then the image reconstruction. While the image-based approach streamlines integration with powerful vision models, it introduces a fundamental challenge: translating reconstructed images back into precise numerical time series values. Although each pixel encodes normalized line graph segments, pixel intensities alone do not inherently carry the original numeric information. Without explicit reference structures or metadata linking pixel values to exact data magnitudes, attempts at reverse engineering the forecasts into their original numeric form prove impractical.

This realization underscores the need for more robust encoding strategies. Future work might embed calibration markers, store minimal numeric anchors, or employ encoding schemes that ensure the one-to-one recoverability of numeric values. Such enhancements would safeguard against information loss and elevate the utility of image-based representations in operational forecasting contexts.

In summary, transforming time series into image-like formats opens the door to powerful vision-based inference models and richer latent representations. This is a novel approach that tackles multivariate TSF tasks with difficulties in input data processing, and reconstruction in both images and numerical data. However, to fully capitalize on this potential, careful attention must be given to how data is encoded and how it can be faithfully retrieved beyond the pixel grid. Hence, we aim to refine this work in the future to mitigate the challenges we face and build a more efficient model.

6. Conclution

We propose a multivariate time series forecasting technique using an image reconstruction approach. We transform each variable of a multivariate time series data into a line graph, stacking them in an image grid, and eventually learning their latent representation using VAE to reconstruct the rightmost 20% of that image following a left-to-right prediction in a traditional TSF task. This approach provides a unique way to tackle multivariate TSF using the capabilities of a more advanced and faster computer vision approach. Experimental results suggest that the selected VAE model showed better results in terms of perceptual and structural similarity compared to the MAE model where its performance was almost similar when measured with pixel-level similarity. Though the MAE had better pixel-level reconstruction results, it did not have a meaningful reconstruction in terms of structural and signal-level information.

References

[1] Performance measurement system (pems) data source. https://dot.ca.gov/programs/traffic-operations/mpr/pems-source. accessed: 2024-09-12. 2, 3

[2] Vitmae. https://huggingface.co/docs/transformers/model_doc/vit_mae. accessed: 2024-09-12. 4

[3] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024. 1, 2, 3, 7

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 2

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2

[7] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997. 1, 2

[8] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[9] 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, 2024. 2, 7

[10] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48(2):210–217, 2020. 2

[11] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71(599-607):6, 1986. 1

[12] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 7

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[14] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1, 2

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5