An efficient ViT-based spatial interpolation learner for field reconstruction

Hongwei Fan$^{1,2[0000-0003-0294-8869]}$, Sibo Cheng$^{1[0000-0002-8707-2589]}$, Audrey J de Nazelle$^{2[0000-0002-1092-3971]}$, and Rossella Arcucci$^{1,3[0000-0002-9471-0585]}$

[1] Data Science Institute, Imperial College London, UK
[2] Centre for Environmental Policy, Imperial College London, UK
[3] Department of Earth Science and Engineering, Imperial College London, UK

Abstract. In the field of large-scale field reconstruction, Kriging has been a commonly used technique for spatial interpolation at unobserved locations. However, Kriging's effectiveness is often restricted when dealing with non-Gaussian or non-stationary real-world fields, and it can be computationally expensive. On the other hand, supervised deep learning models can potentially address these limitations by capturing underlying patterns between observations and corresponding fields. In this study, we introduce a novel deep learning model that utilizes vision transformers and autoencoders for large-scale field reconstruction. The new model is named ViTAE. The proposed model is designed specifically for large-scale and complex field reconstruction. Experimental results demonstrate the superiority of ViTAE over Kriging. Additionally, the proposed ViTAE model runs more than 1000 times faster than Kriging, enabling real-time field reconstructions.

Keywords: Field Reconstruction · Vision Transformer · Deep Learning.

1 Introduction

Spatial interpolation, which is predicting values of a spatial process in unmonitored areas from local observations, is a major challenge in spatio-temporal statistics. As a reference method of spatial interpolation, Kriging [18,5] provides the best linear unbiased prediction from observations. As a Gaussian process [21] governed by covariance, Kriging interpolates unmonitored areas as a weighted average of observed data. Kriging [18,5] is a geostatistical method that provides the optimal linear unbiased prediction based on observed data. It assumes that the underlying data follows a Gaussian process and is governed by covariance. However, authors [19,26] noted that in many cases, the spatial covariance function of physical fields is non-Gaussian and non-stationary. As a consequence, the optimality of Kriging can not be guaranteed in real-world scenarios. Another limitation of Kriging is that its computational complexity can render it impractical for large spatial datasets. In fact, the online implementation of Kriging involves computing the inversion of a $N \times N$ covariance matrix, where N is the number of observed locations [13]. The aforementioned limitations of Kriging pose

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_34

significant challenges in utilizing this method for generating credible large-field reconstructions in real-time.

Recently, deep learning (DL) [15] or neural network (NN) [14] have become increasingly utilized and powerful prediction tools for a wide range of applications[11], especially in computer vision and natural language processing [15]. DL has witnessed an explosion of architectures that continuously grow in capacity [22]. The utilization of convolutional neural network (CNN)s has become increasingly prevalent due to the rapid progress in hardware. CNNs are well-suited for prediction tasks that involve complex features, such as nonlinearity and non-stationarity, and offer computational efficiency when analyzing massive datasets with GPU acceleration. Previous research efforts, such as [23], have explored the use of DL techniques for field reconstructions from observations. However, the traditional CNN-based approaches are inadequate in dealing with the problem of time-varying sensor placement. Re-training is often required when the number or the locations of sensors change, resulting in difficulties of real-time field construction. Therefore, this paper proposes the use of DL for large field reconstruction from random observations in real time.

2 Related works and contribution of the present work

NNs [15] have become a promising approach for effectively reconstructing fields from sparse measurements. [8,1,25]. While graph neural network (GCN) [24] and multi-layer perception (MLP) [15] could handle sparse data, they are known to scale poorly because of the high computational cost. Moreover, these methodologies require predetermined measurements as input data, which renders them unfeasible for real-world situations where sensor quantities and positions frequently vary over time, ultimately making them impractical [7]. To tackle these two bottlenecks, Fukami et al. [9] utilized Voronoi tessellation to transfer observations to a structured grid representation, which is available for CNN. Despite their effectiveness in capturing features, CNN typically overlook the spatial relationships between different features, thus making it difficult for them to accurately model the spatial dependencies required for reconstructing large-scale fields [16]. This difficulty can be addressed by the introduction of Vision Transformers (ViT) [6]. Transformers were proposed by Vaswani et al. (2017) [22] and have since become the state-of-the-art method in machine translation [20]. Apart from the complex architecture of transformers, the effectiveness of natural language processing (NLP) models heavily relies on the training strategy employed. One of the critical techniques utilized in training is auto-encoding with masks [12], where a subset of data is removed, and the model learns to predict the missing content. This technique has also demonstrated promising results in the field of computer vision, further highlighting its potential for enhancing model performance. By using masked autoencoder (AE)s to drop random patches of the input image and reconstruct missing patches, He et al. [12] demonstrates that it is possible to reconstruct images that appear realistic even when more than 90% of the original pixels are masked. The underlying principle of reconstructing an image from

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_34

randomly selected patches bears resemblance to the process of reconstructing a field from observations.

Although the ViT model and the AE method [12] have succeeded in image reconstruction, to the best of our knowledge, no previous studies have applied them to field reconstruction task. Inspired by the success of ViT and AE methods, we propose a simple, effective, and scalable form of a Vision Transformer-based autoencoder (ViTAE) for field construction. To address the challenges mentioned earlier, we present a technique ViTAE that incorporates sparse sensor data into a Transformer model by mapping the observed values onto the field grid and masking unobserved areas. The masked observations field is divided into patches and fed into the transformer encoder to obtain representations. These representations are reshaped into patches and concatenated before being fed into the decoder to predict the grid values. Our proposed model, ViTAE, is capable of efficiently and accurately reconstructing fields from unstructured and time-varying observations. We compare the performance of our ViTAE with Kriging-based field reconstruction methods in this study.

The rest of the paper is organized as follow. Section 3 introduces the construction and properties of our ViTAE method. Section 4 presents some studies to show the performance of ViTAE. Section 5 summarizes our main results and suggests directions for future work.

3 Methodology

Our objective is to reconstruct a two-dimensional global field variable Q ($\dim(Q) = n$) from some local observations $\{\tilde{Q}_i\}, i \in O$ where O is a subset of $[1, ..., n]$. The proposed ViTAE is an autoencoder that aims to reconstruct the complete field from its partial observations. To deal with the sparsity of the data and extract meaningful representations, we employ a ViT-based autoencoder, which enables us to process the observations. Figure 1 illustrates the flowchart of the proposed approach.

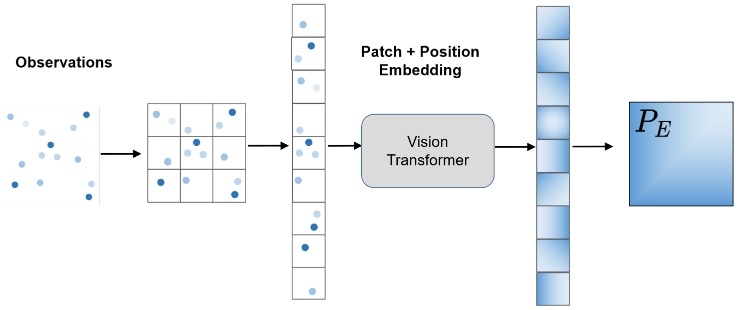

Fig. 1. Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder.

DOI: 10.1007/978-3-031-36027-5_34

ViT-based autoencoder First, to allow computationally tractable use of ViT, we project local sensor measurements into a grid field $\boldsymbol{I}$ according to their location in the field, defined as:

$$\boldsymbol{I}_i\,(i = 1, ..., n) = \begin{cases} \tilde{Q}_i & \text{if } \tilde{Q}_i \text{ is observable (i.e., } i \in O) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Just as in a standard ViT, we reshape the field $\boldsymbol{I}$ into a sequence of N flattened 2D patches, N is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer uses a constant latent vector size D through all of its layers, so we flatten the patches and map to the embedding of dimensions D by a linear projection with added positional embeddings, and then process the resulting set via a series of Transformer blocks to obtain latent representations and map the latent representation back to the predicted field P_E.

Reconstruction target Our ViTAE reconstructs the field by predicting the values for entire field grids. Unlike [12] which calculates the loss only on the masked patch, we compute the loss in the entire field. The loss is defined as:

$$L = MSE\,(Q, P_E) \tag{2}$$

it computes the mean squared error (MSE) between the reconstructed P_E and original fields Q in pixel space.

4 Test cases and results

This section presents an evaluation of the performance of ViTAE compared to Kriging in stationary simulation data with a structured grid. To create a spatially isotropic physical field, we generated the simulation field data using the Gaussian covariance kernel with the help of gstool [17]. As a result, the correlation between two points in the field is dependent solely on their spatial distance, which is ideal for the Kriging method. Such simulations are commonly employed for comparing various field reconstruction approaches [2]. After generating the grid field, a random selection of grid points is used as observations for field reconstruction. Unlike the ViTAE method, Kriging requires prior knowledge of the covariance kernel, including the kernel function and correlation length. To further investigate the robustness of the method, numerical experiments of Kriging are conducted using two kernel functions: Gaussian and Exponential, both with the same correlation length as used for data generation. The latter is done to simulate cases where the kernel function is misspecified, as in real-world applications, the exact covariance kernel is often unknown [10].

Our initial focus is on the computational efficiency of the proposed method. Table 1 displays the accuracy of the reconstructing fields using the Gaussian kernel of different sizes with varying numbers of observations. The results indicate

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_34

that when the field size is larger than 256 and the number of observations exceeds 0.1%, Kriging's computational time grows exponentially, taking thousands of seconds to fit and predict. It should be noted that Kriging must be performed online, i.e., after the observations are available, which poses computational challenges for high-dimensional systems. To compare the reconstruction accuracy of our proposed ViTAE against Kriging, we conducted experiments on a field size of 512×512. We used 0.5%, 1%, 2%, and 5% of the total number of grid points in the field as the number of observations for training. For each observation ratio, we generated 10,000 field snapshots and randomly selected observations from each snapshot. This allowed us to use time-varying observations as input data for Kriging and ViTAE to learn the entire physical field. We randomly partitioned our dataset into training, validation, and testing sets using an 80/10/10 split.

Model	ϵ			
Kriging/RBF	0.2243	0.2221	0.2218	0.2215
Kriging/Exp	0.2553	0.2552	0.2550	0.2379
ViTAE-lite/16	0.2431	0.2346	0.2290	0.2242
ViTAE-base/16	0.2280	0.2369	0.2250	0.2234
ViTAE-large/16	**0.2255**	**0.2228**	**0.2213**	0.2202
Sampling Percent	0.5%	1%	2%	5%

Table 1. Gaussian field reconstruction result of the Gaussian field reconstruction for ViTAE, and Kriging.

Model variation For the ViT-based encoder design, we follow the original ViT set up [6] and use "Lite", "Base", and "Large" models, such that ViT-Lite has 8 layers, 32 as hidden size, 8 as Heads, 16 hannels and 16 as Patch size, ViT-Base has 8 layers, 64 as hidden size, 32 as Heards, 16 hannels and 16 as Patch size, and ViT-Large has 8 layers, 128 as hidden size, 64 as Heads, 16 hannels and 16 as Patch size. In what follows we use brief notation to indicate the model size and the input patch size: for instance, ViT-L/16 denotes the "Large" variant with 16×16 input patch size.

The field reconstruction results are shown in Figure 2. The field reconstructed by ViTAE shows great similarity against the ground truth (GT) without knowing the spatial correlation function *a priori*.

Figure 2 also reports the relative error defined as:

$$\epsilon = \frac{\|Q_{\text{ref}} - Q_{\text{reconstruct}}\|_2}{\|Q_{\text{ref}}\|_2}, \tag{3}$$

where $\|\cdot\|$ denotes the L_2 norm, and Q_{ref} and $Q_{\text{reconstruct}}$ are the reference and reconstructed simulation fields, respectively. This metric of relative error has been widely used in field reconstruction and prediction tasks [4,3]. In this section,

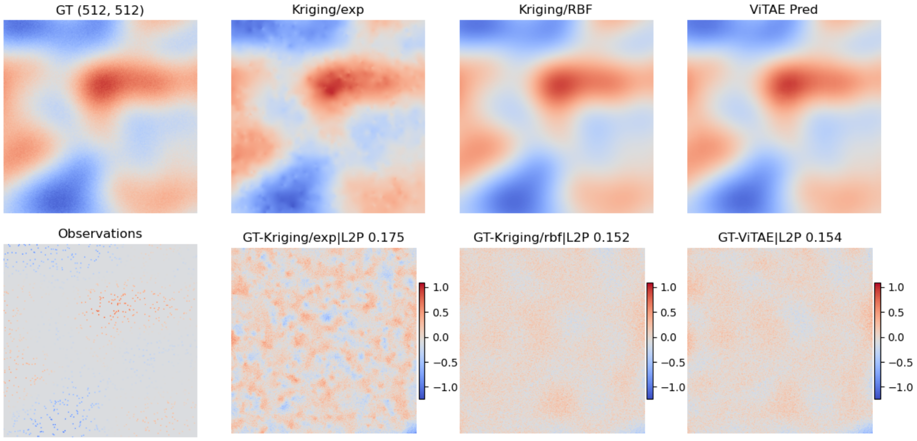

Fig. 2. 512 × 512 Gaussian field reconstruction results of ViTAE and Kriging , 0.5% sampling rate comparet to the GT.

we compare the performance of our ViTAE model with two Kriging models using Gaussian and exponential covariance kernels, denoted as Kriging/RBF and Kriging/exp, respectively. As shown in Figure 2, Kriging/exp is significantly outperformed by Kriging/RBF, which demonstrates the vulnerability of Kriging when the covariance kernel is not perfectly known. On the other hand, our ViTAE model, which does not require prior knowledge of the covariance kernel, achieves reconstruction results that are almost as accurate as Kriging/RBF. Additionally, the online computation of ViTAE is much more efficient than Kriging, as shown in Table 2. For example, when 5% of the field is observable, ViTAE-lite/16 runs 10^6 faster than Kriging.

Model	Execution time (s)			
Kriging/RBF	21	59	191	1491
Kriging/Exp	31	76	253	1586
ViTAE-lite/16	0.0105	0.0104	0.0105	0.0106
ViTAE-base/16	0.0128	0.0127	0.0128	0.0128
ViTAE-large/16	0.0150	0.0154	0.0151	0.0153
Sampling Percent	0.5%	1%	2%	5%

Table 2. Execution time in seconds of the Gaussian field reconstruction for ViTAE, and Kriging.

DOI: 10.1007/978-3-031-36027-5_34

5 Conclusion

A long-standing challenge in engineering and sciences has been spatial interpolation for large field reconstruction. To tackle this issue, the paper introduces a novel autoencoder based on the ViT architecture, which serves as an efficient learner for spatial interpolation. The results presented in this paper indicate that the proposed ViTAE approach outperforms the Kriging method in spatial interpolation tasks. The method does not require prior knowledge of the spatial distribution and is computationally efficient. This work opens up new possibilities for applying DL to spatial prediction and has potential applications in complex data structures. In addition, the proposed method can be extended to real-world physical systems to investigate relationships and correlations between observations, supporting studies in spatio-temporal statistics and geostatistics.

References

1. Bolton, T., Zanna, L.: Applications of deep learning to ocean data inference and subgrid parameterization. Journal of Advances in Modeling Earth Systems **11**(1), 376–399 (2019)
2. Chen, W., Li, Y., Reich, B.J., Sun, Y.: Deepkriging: Spatially dependent deep neural networks for spatial prediction. arXiv preprint arXiv:2007.11972 (2020)
3. Cheng, S., Chen, J., Anastasiou, C., Angeli, P., Matar, O.K., Guo, Y.K., Pain, C.C., Arcucci, R.: Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models. Journal of Scientific Computing **94**(1), 1–37 (2023)
4. Cheng, S., Prentice, I.C., Huang, Y., Jin, Y., Guo, Y.K., Arcucci, R.: Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. Journal of Computational Physics p. 111302 (2022)
5. Cressie, N.: The origins of kriging. Mathematical Geology **22**, 239–252 (1990)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Environmental Research Group, I.C.L.: London air quality network. `https://www.londonair.org.uk/LondonAir/Default.aspx` (2022 (accessed Nov 8, 2022))
8. Erichson, N.B., Mathelin, L., Yao, Z., Brunton, S.L., Mahoney, M.W., Kutz, J.N.: Shallow neural networks for fluid flow reconstruction with limited sensors. Proceedings of the Royal Society A **476**(2238), 20200097 (2020)
9. Fukami, K., Maulik, R., Ramachandra, N., Fukagata, K., Taira, K.: Global field reconstruction from sparse sensors with voronoi tessellation-assisted deep learning. Nature Machine Intelligence **3**(11), 945–951 (2021)
10. Ginsbourger, D., Dupuy, D., Badea, A., Carraro, L., Roustant, O.: A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. Applied Stochastic Models in Business and Industry **25**(2), 115–131 (2009)
11. Hadash, G., Kermany, E., Carmeli, B., Lavi, O., Kour, G., Jacovi, A.: Estimate and replace: A novel approach to integrating deep neural networks with existing applications. arXiv preprint arXiv:1804.09028 (2018)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_34

12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
13. Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., et al.: A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics **24**(3), 398–425 (2019)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
16. Linsley, D., Kim, J., Veerabadran, V., Windolf, C., Serre, T.: Learning long-range spatial dependencies with horizontal gated recurrent units. Advances in neural information processing systems **31** (2018)
17. Müller, S.: Geostat framework. `https://geostat-framework.org/` (2022 (accessed Nov 8, 2022))
18. Oliver, M.A., Webster, R.: Kriging: a method of interpolation for geographical information systems. Int. J. Geogr. Inf. Sci. **4**, 313–332 (1990)
19. Paciorek, C.J., Schervish, M.J.: Spatial modelling using a new class of nonstationary covariance functions. Environmetrics: The official journal of the International Environmetrics Society **17**(5), 483–506 (2006)
20. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
21. Rasmussen, C.E.: Gaussian processes in machine learning. In: Summer school on machine learning. pp. 63–71. Springer (2003)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on epi. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6319–6327 (2017)
24. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems **32**(1), 4–24 (2020)
25. Yu, J., Hesthaven, J.S.: Flowfield reconstruction method using artificial neural network. Aiaa Journal **57**(2), 482–498 (2019)
26. Zareifard, H., Khaledi, M.J.: Non-gaussian modeling of spatial data using scale mixing of a unified skew gaussian process. Journal of Multivariate Analysis **114**, 16–28 (2013)