Acta Geotechnica (2025) 20:4977–5005
https://doi.org/10.1007/s11440-025-02690-z

RESEARCH PAPER

Correlation of excavated soil multi-source heterogeneous data using multimodal diffusion model

Qi-Meng Guo[1,2,3] · Liang-Tong Zhan[2,3] · Zhen-Yu Yin[1,4] · Hang Feng[1] · Guang-Qian Yang[5] · Yun-Min Chen[2,3] · Yu-An Chen[2,3]

Received: 31 December 2024 / Accepted: 2 June 2025 / Published online: 21 July 2025

Abstract

The sustainable utilization of excavated soil as a geomaterial requires a comprehensive understanding of its multi-dimensional properties, but correlating heterogeneous data (e.g., visual, mechanical, and electrical characteristics) remains a challenge. To address this, an excavated soil information collecting system was developed to acquire multi-source data including RGB images, cone index (CI) curves, and TDR waveforms—from China's largest soil transfer platform, establishing a database of 23,122 sets. A generative-model-aided correlation analysis framework was proposed, leveraging a denoising diffusion probabilistic model to explore inherent relationships between soil properties. Performance metrics, such as SSIM, LPIPS, and RMSE, were employed to analyze the model's training results. Key findings reveal that: (1) soil images encode water content information, which correlates with CI curves and TDR waveforms; (2) CI and TDR data cannot capture color-based mineral composition details from images; and (3) TDR waveforms uniquely detect pollution indicators (e.g., electrical conductivity), undetectable via other methods. This AI-driven approach provides a novel methodology for analyzing multi-dimensional property correlations in geotechnics, enhancing sustainable soil reuse.

Keywords Denoising diffusion probabilistic model · Excavated soil · Generative model · Inherent correlation · Multi-source heterogeneous data

✉ Liang-Tong Zhan
zhanlt@zju.edu.cn

✉ Zhen-Yu Yin
zhenyu.yin@polyu.edu.hk

Qi-Meng Guo
qimeng.guo@connect.polyu.hk; qimengguo@zju.edu.cn

Hang Feng
fenghang.feng@connect.polyu.hk

Guang-Qian Yang
guangqian.yang@connect.polyu.hk

Yun-Min Chen
chenyunmin@zju.edu.cn

Yu-An Chen
chen-yu-an@zju.edu.cn

[1] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[2] MOE Key Laboratory of Soft Soils and Geoenvironmental Engineering, Zhejiang University, Hangzhou 310058, China

[3] Institute of Geotechnical Engineering, Zhejiang University, Hangzhou 310058, China

[4] Research Centre for Resources Engineering Towards Carbon Neutrality (RCRE), The Hong Kong Polytechnic University, Kowloon, Hong Kong

[5] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

1 Introduction

1.1 Excavated soils: sustainable geomaterials with potentials for utilization

Excavated soils are geomaterials generated from building foundation pit excavation, tunnel excavation, channel excavation, and other engineering excavations [23]. In developed countries like European Union, excavated soils account for 59% of the total construction and demolition wastes (CDW) generation [16]. In developing countries like China, excavated soils even account for 70–80% of CDW generation [22]. Excavated soils are currently regarded as waste under EU law and commonly disposed in landfills [18, 21]. However, more than 90% of these soils are not contaminated and could be utilized according to their engineering properties, environmental properties, and resource potentials [56]. For example, pure soils with moisture content below 1.5 times liquid limit can be directly backfilled as agricultural land or construction land; inert soils with organic matter content less than 5% can be compacted-solidified or casted-solidified as engineering fillers; soils with pH value 5–10, moisture content $< 40\%$, and LOI $< 50\%$ can be used as sinter-free products or sintered products [57]. Obviously, the effective reuse of environment-friendly excavated soils as sustainable geomaterials depend on a comprehensive understanding of multiple characteristics including sand content, organic matter content, mineral compositions, and so on. These properties are typically captured through diverse detection methods—such as image-based analysis, stress–strain curve fitting, and electrical waveform profiling—yet their interdependencies remain underexplored in practice. The identification of interrelationships among diverse indicators facilitates the use of easily measurable parameters as proxies for critical but labor-intensive measurements, allowing for efficient on-site assessment and classification of excavated soils while improving sustainable geomaterial management.

1.2 Inherent correlations of soil multi-dimensional properties

Studying the correlations between soil properties is a classic topic in geomaterial research, helping to infer indicators that are difficult to measure directly based on easily obtainable ones. Carter and Bentley [11] presented typical values of engineering properties for various types of soils, together with correlations between different properties. By analyzing a large amount of laboratory data, correlations between difficult-to-measure indices (e.g., frost susceptibility and swelling potential) and classical indices (e.g., grading, plasticity, density, permeability, consolidation, and shear strength) were discussed [12]. Furthermore, Ameratunga et al. [3] combined numerous field data from in situ testing such as standard penetration test (SPT), cone penetration test (CPT), pressuremeter test (PMT), dilatometer test (DMT), and vane shear test (VST), bridging the gap between laboratory data and field data. Verbrugge and Schroeder [52] compiled empirical formulas between soil properties and classified them in different scenarios from the perspective of geotechnical practitioners.

Nowadays, besides studying soil engineering properties [26, 59], some scholars have begun to consider environmental properties and resource properties, and comprehensively analyze the soil multi-source heterogeneous data [17, 46]. Meimaroglou and Mouzakis [36] investigated the influence of clay fraction content, specific surface area (SSA), cation exchange capacity (CEC), and mineralogy on earth mortars. It was found that compressive strength depends mainly on SSA and CEC, with weaker correlations for clay content and iron oxides. Sainju and Liptzin [44] and Sainju et al. [45] related soil physicochemical properties (soil pH, electrical conductivity, CEC, and nutrient concentrations) to soil health properties. The importance of CEC, inorganic P, and K was identified, and some novel indicators such as average slake aggregate (ASA) was proposed. Studies about correlations between soil physicochemical properties and biological properties were also implemented. Soil organic matter content (SOM), particle size distribution (PSD), CEC, porosity, and water holding capacity are all critical indices discussed [28, 34, 51]. In addition, soil images containing texture and color information were noticed. Teixeira and Basch [49] carried out a campaign to explore the correlations between visual soil assessment (VSA) soil indicators (e.g., structure, porosity, stability, soil color, and surface ponding) and measured soil properties (infiltration rate, pH and labile organic carbon, SOM). Olivares et al. [39] analyzed the relationships between the visual evaluation of soil structure (VESS) and soil properties. VESS is validated to be a reliable semi-quantitative method to assess soil quality and could be considered a promising visual predictor of soil physical properties such as bulk density, SOM, and soil penetration resistance.

To note, different soil indicators are multi-source heterogeneous. For example, the result of CPT is a curve distributed along the depth, and pH is just a value. Therefore, when exploring the correlation of these indicators, it is necessary to artificially adjust high-dimension data, such as extracting the uniformity coefficient (C_u) and coefficient of curvature (C_c) from the PSD curve to get some single values for direct analysis, but this operation destroyed the richness of the raw data. Classical correlation

analysis methods include Pearson r correlation, Kendall rank correlation, and Spearman rank correlation. Pearson r correlation excels at measuring linear relationships between two continuous variables, providing intuitive results (-1 to 1) with strong interpretability. Spearman rank correlation is suitable for assessing monotonic relationships (linear or nonlinear) based on data ranking, while Kendall rank correlation measures ordinal concordance, making it ideal for small samples or data with numerous ties—its results are equally intuitive ($\tau = 1$ indicates perfect agreement, $\tau = -1$ perfect disagreement). Both Spearman and Kendall methods can analyze sequential data correlations [9, 10, 13]. Fan et al. [19] investigated aged/rejuvenated asphalt's chemical-rheological correlations through DSR, BBR, and FTIR tests. Gray entropy correlation analysis (GECA) revealed strong correlations (> 0.96) between SI/CI indices and rheological parameters (G^*, δ, S). Ching et al. [15] developed a multivariate probability distribution model for coarse-grained soil parameters using the SAND/7/2794 database. The model effectively captures parameter correlations and serves as a Bayesian prior, updatable with site-specific data. While offering a consistent uncertainty integration framework, the authors caution against extrapolation beyond the database scope, noting similar limitations apply to conventional regression approaches. Roy et al. [43] employed machine learning (GPR, RFR, DTR) to analyze correlations between mix design parameters and mechanical properties of rice husk ash concrete. The high R^2 values (0.964–0.969) demonstrate strong predictive correlations, particularly for DTR models. PDP analysis further reveals key parameter-strength relationships, providing quantitative correlation insights that surpass conventional experimental approaches. The ML framework effectively captures nonlinear correlations, offering a robust alternative to labor-intensive laboratory testing for RHAC optimization. Liu et al. [32] developed novel correlations between resilient modulus (M_r) and CPTU indices for clayey soils using a multivariate normal distribution framework. By combining Box-Cox transformation, Pearson correlation analysis, and Bayesian updating with bootstrap uncertainty quantification, the method reliably predicted M_r from cone tip resistance, sleeve frictional resistance, moisture, and dry density. While showing good accuracy for Jiangsu clays (124 datasets), caution is needed for global applications due to potential regional biases. In recent years, with the advancement of deep learning, neural networks have gained popularity in correlation analysis. They are applicable not only to bivariate correlations but also to ternary and more complex relationships [58]. Michon et al. [38] employed quantitative structure–property relationship (QSPR) methods and neural networks to identify nonlinear relationships between chemical and rheological properties. Strechan et al. [48] used artificial neural networks (ANNs) to derive correlations between the enthalpy of vaporization, the surface tension, the molar volume, and the molar mass of a substance. Karabulut and Koyuncu [27] developed neural network models to establish correlations of thermal conductivity with temperature and density for propane. Asghari et al. [4] proposed a DNN-based framework for analyzing complex correlations in engineering metrics, utilizing 1,101 clay samples to study the relationship between undrained shear strength and factors such as liquid limit, plastic limit, water content, vertical effective stress, and preconsolidation stress, demonstrating strong performance in handling nonlinear interactions and uncertainties. Alessandrini et al. [2] introduced a neural network-based correlation analysis framework to enhance electroencephalography (EEG)-speech stimulus response detection. By implementing a single multilayer perceptron (MLP) with a correlation-optimized loss function, their method outperformed traditional linear canonical correlation analysis (CCA), achieving a 10.56% improvement in Pearson correlation. While these studies primarily focus on numerical (0D) and some sequential (1D) data, they have yet to explore correlations involving higher-dimensional data such as images (2D).

1.3 Multimodal generative models for exploring soil properties

With the popularity of generative AI platforms like ChatGPT, Google Bard, DALL-E, and Musico, the potential of generative models including variational autoencoder (VAE), generative adversarial network (GAN), and diffusion model in geotechnics has begun to gain attention [8, 47]. As for soil data optimization, Bai et al. [7] developed an improved super-resolution method, SRLGAN, for reconstructing high-resolution soil CT images, addressing limitations like blurred boundaries and low quality. Meng et al. [37] proposed the SS GAN, an unsupervised shadow removal algorithm for soil surface images, to improve soil moisture content estimation accuracy. Wang et al. [53] proposed GCS-CVAE to address missing data and high energy consumption in wireless soil sensors. GCS-CVAE demonstrated superior reconstruction accuracy, stability, and efficiency in soil monitoring data. As for soil behavior prediction, Tsimpouris et al. [50] proposed a novel stacked autoencoder-based methodology for transforming soil spectra into a compressed latent space to improve soil property prediction accuracy. Applied to LUCAS 2009 topsoil data, it reduced RMSE by up to 9.9%, enabling simultaneous prediction of PSD, pH, CEC, organic carbon, calcium carbonate, and total nitrogen. He et al. [24] proposed a dynamic SOM estimation model using GAN to enhance hyperspectral datasets. By

generating pseudo-samples, the best model, GAN-BPNN, achieved a 30.8% R^2 increase and 44.5% RMSE reduction. Lo Man et al. [33] used VAE to predict embankment settlement and pore water pressure directly from monitoring data, eliminating the need to update soil parameters. Chen et al. [14] introduced a hybrid deep learning model combining CVAE and Kriging to predict soil properties from sparse geotechnical data, which was applied to CPT results. Obviously, generative models are showing great potential in geotechnical data imputation and property prediction due to the capacity of decoding relationship between different forms of data.

Geotechnical researchers usually use classic methods such as Spearman correlation analysis, Pearson correlation analysis, and principal component analysis when studying data correlations [34, 49, 51]. These methods are good at dealing with property correlation research for data of the same dimension but cannot handle complex correlation research involving multi-source heterogeneous data such as images, waveforms, and values. The Pearson/Spearman correlation coefficients require one-dimensional numerical vectors as input and thus cannot directly process two- or three-dimensional data (e.g., images). Flattening an image into pixel vectors (e.g., converting 100 × 100 pixels to a 10,000-dimensional vector) for correlation analysis with waveform (e.g., 1000 data points) would lead to dimensional mismatch (necessitating forced alignment that compromises physical meaning) and numerical sensitivity (where unit difference between pixel RGB values and voltage measurements may induce spurious correlations). While PCA is partially applicable for cross-modal correlation analysis, it demands rigorous data preprocessing—requiring transformation of multimodal data into feature vectors with consistent dimensions. This process risks losing critical modal characteristics (e.g., local image topology or waveform phase information), and the results become highly sensitive to subjective feature selection decisions. Conversely, generative models are experts in grasping the physical patterns behind real data, and generate new data based on the learned physical laws, which implies potentials for correlation research. Generative models can directly process raw heterogeneous inputs—such as 2D/3D images, 1D waveforms, and 0D numerical values—while preserving the complete information of each modality. By leveraging the multilayer architecture of neural networks to approximate arbitrarily complex functions and activation functions to introduce nonlinearity, these models enable deep mining of nonlinear relationships across modalities. Therefore, how to use advanced generative AI tools to explore multi-source heterogeneous soil properties has become the focus of researchers.

1.4 Objective of the study

In the present work, a generative-model-based analysis framework for soil multi-source heterogeneous properties was proposed. Excavated soils with significant differences in engineering properties, environmental properties, and resource potentials were regarded as the studied geomaterials. An excavated soil information collecting system (ESICS) was developed to collect soil multi-source heterogeneous data including soil images (2D information), cone index curves (1D spatial-series data), and TDR waveforms (1D time-series data). The advanced generative model, denoising diffusion model, -aided investigation on inherent correlations of soil multi-source heterogeneous data was carried out, including three cases of data conversion and generation (soil image-cone index, soil image-TDR waveform, and cone index-TDR waveform). Some inherent relationship between soil imaging information, mechanical properties, and electrical properties was found. This study provides a novel perspective with the support of artificial intelligence methods for exploring the correlation between geomaterial properties.

2 Principle and methodology

2.1 Soil properties and their correlations

Figure 1 shows soil three types of properties and their triangle correlations. The engineering properties include moisture content, grading, plasticity, permeability, consolidation, and shear strength, which are classical concepts in geotechnics. The environmental properties include pH, electrical conductivity, CEC, dissolved inorganic compounds (e.g., sulfate, nitrate), and heavy metal content, which are key indicators in environmental science. The resource properties reflect the potential for utilization of the geomaterials and typically include chemical elements, mineralogy, SOM, and nutrients (e.g., nitrogen, phosphorus, potassium), which are of interest in soil science and agronomy. These properties are intrinsically interrelated and determine the sustainable utilizations of geomaterials. For example, as for engineering attribute affecting environmental attribute, geomaterials with more fine content tend to adsorb more heavy metal-like pollutants; as for environmental attribute affecting engineering attribute, geomaterials with pollutants after biochemical degradation will lose mass and have changes in grading, leading to different constitutive behaviors. Regarding engineering attribute affecting resource potential, low compressibility geomaterials dominated by coarse sands are likely to serve as aggregates for construction materials; regarding

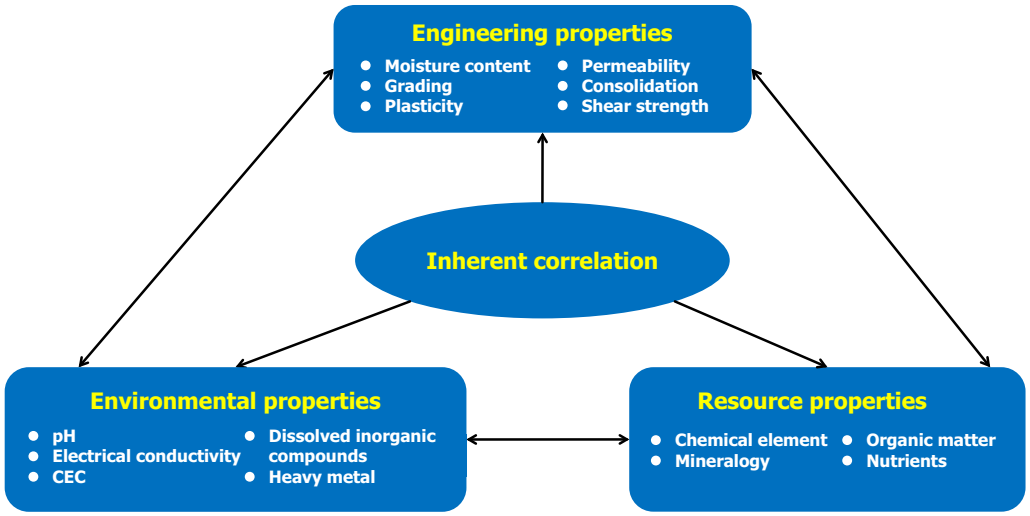


Fig. 1 Soil properties and their inherent correlations

resource potential affecting engineering attribute, geomaterials with rich montmorillonite content are capable of absorbing more water. For environmental attribute affecting resource potential, acid soil or alkaline soil can serve as planting materials for certain crops; for resource potential affecting environmental attribute, peat soils with excessive SOM tend to adsorb more heavy metals.

For instance, in image analysis applications, soil images contain both engineering and resource-related property information. Paul et al. [41] analyzed mortar slump variations under different rice husk powder ratios through macroscopic mortar images while examining microstructural composition via SEM. Similarly, Kashyap et al. [29] used SEM to observe hydration reactions in modified concrete, explaining strength variations microscopically. At macroscopic scales, deep learning has emerged as a primary tool for geomaterial identification—Zhao et al. [60] developed an enhanced YOLO model for automated mucky soil classification in tunneling, while Yan et al. [54] created a Bayesian-optimized image augmentation framework with AlexNet/GoogLeNet for improved muck identification. Therefore, establishing bidirectional correlations between soil images and mechanical/electrical properties (either predicting properties from images or generating visual characteristics from physical parameters) would enhance understanding of fundamental physical mechanisms while enabling interpretable prediction of difficult-to-measure properties from easily obtainable indicators.

Soil properties are typically represented in different forms of data. Classical correlation analysis methods, such as Spearman correlation analysis, Pearson correlation analysis, and principal component analysis, are applicable to analyzing the relationships between values (0D). However, due to the diversity of detection methods, 1D, 2D, and even 3D indicators are more common as raw data in testing. In order to analyze these data with different dimensions, conventional analysis methods often extract critical information from high-dimension data. This dimensionality reduction process frequently involves information loss, making it difficult to fully characterize the true properties of the target medium. Consequently, the identification of inherent correlations between properties may be biased or overlooked. This study aims to explore nondestructive correlation analysis methods for multi-dimension data by using soil images (2D data), cone index curves (1D spatial-series data), and TDR waveforms (1D time-series data).

2.2 Typical generative models

In recent years, with the rapid development of deep learning, deep generative models have been highly favored as a novel data processing tool across various disciplines [1, 35]. Among them, the most commonly used and efficient approaches are variational autoencoders (VAEs), generative adversarial networks (GANs), and lastly diffusion models (DMs). Figure 2 illustrates their mechanisms.

VAEs, introduced by Kingma [30], are probabilistic generative models combining autoencoders and variational inference. VAEs encode data into a lower-dimensional latent space with a probabilistic interpretation. These low-dimensional latent variables (z) are latent representation of the input data, meaning they are abstract symbols of useful information and physical laws. Later, VAEs decode the latent space and enable both data reconstruction and new sample generation by sampling from the latent space. Usually, the probability distribution of these latent variables is denoted by Gaussian distribution. VAEs have been

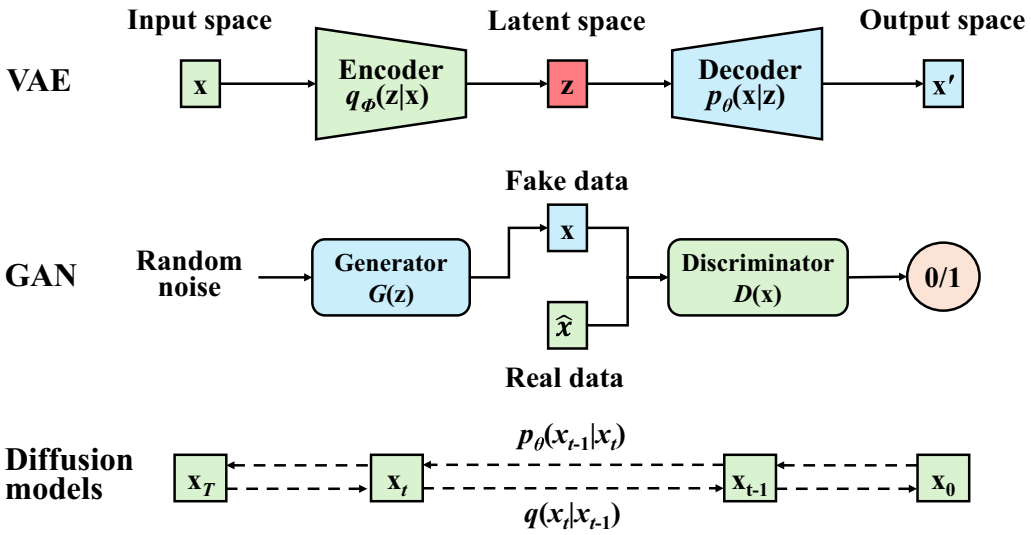


Fig. 2 Typical generative models

applied to tasks like image synthesis, anomaly detection, and representation learning.

GANs, introduced by Goodfellow et al. [20], revolutionized generative models with adversarial training framework. GANs consist of two neural networks: a generator and a discriminator. The generator learns to produce realistic data samples, while the discriminator distinguishes between real and fake data. These networks engage in a minimax game: the generator aims to "fool" the discriminator by generating realistic data, while the discriminator improves its ability to identify fake samples. Through this adversarial process, both networks enhance each other iteratively, enabling GANs to achieve cutting-edge performance in tasks like image synthesis and data generation. GANs were considered state-of-the-art generative models until the recent rise of DMs.

DMs, i.e., denoising diffusion probabilistic models (DDPMs) introduced by Ho et al. [25], are generative models that create data by reversing a gradual noising process. The framework involves two steps: a forward process that adds random noise to input data step-by-step, transforming it into pure Gaussian noise, and a reverse process that learns to denoise this noise to reconstruct the original data. By using neural networks, the model learns to fine noise distribution added to training data, and the realistic image can be reconstructed by gradually removing the noise. This reverse process is modeled using deep neural networks like U-Nets or transformers. By explicitly modeling the denoising steps, DMs achieve high-quality and diverse sample generation. Unlike GANs, DMs are likelihood-based, providing a stable training process and interpretable generation. While GANs and VAEs have shown success in data generation, their limitations in modeling soil's multimodal nonlinearities motivated our choice of DDPM. Specifically: (1) The adversarial loss of GANs does not have the incentive to cover the entire data distribution. When the discriminator has been over trained or catastrophic forgetting happens, the generator might tend to produce a small part of the data diversity leading to mode collapse; (2) the latent space of VAEs is much smaller than the image. This induces the model to predict an average of pixels to find the optimal solution, resulting in a blurry image. The low-quality generation might lead to the information loss in the process of cross-modal generation.

The evaluation of generative models spans three key points: high-quality samples, mode coverage (diversity), and fast sampling. While GANs excel in high-quality samples and fast sampling, and VAEs achieve good mode coverage with fast sampling, diffusion models (DMs) uniquely ensure both high-quality samples and comprehensive mode coverage. In this study, given that the resolution of augmented image samples is modified to only 84×84, the impact of sampling rate changes exceeding a certain threshold on the quality of generated samples can be negligible. Therefore, putting data diversity and sample fidelity over sampling rate justifies the selection of DMs for investigating inherent correlations between soil multi-source heterogeneous data.

2.3 Correlation analysis framework

Based on the principles and characteristics of generative models, combined with the need for correlation analysis of multi-dimensional multi-source heterogeneous data, a generative-model-based framework for soil property correlation analysis is proposed (Fig. 3). Assuming that index

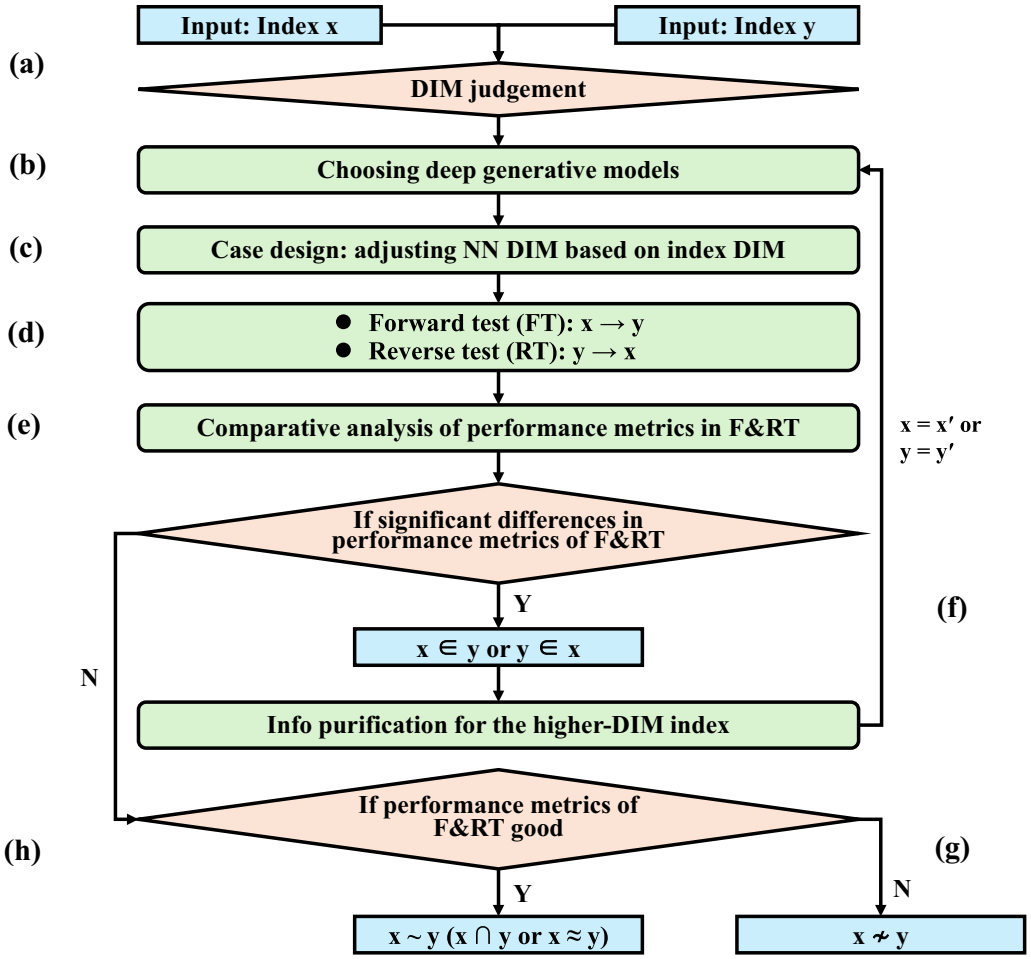


Fig. 3 Generative-model-based correlation analysis framework for soil properties

x is RGB image and index y is one-dimensional waveform, the framework's workflow can be illustrated as follows:

(a) **Determine the dimensions of the input variables x and y:** Identify whether the data is 3D (diorama), 2D (image), 1D (waveform), or 0D (value).

(b) **Select an appropriate deep generative model:** Choose a model such as VAEs, GANs, or DMs based on the data requirements.

(c) **Design the model structure based on data dimensions:** Adjust the convolutional neural network (CNN) architecture in the generative model according to the input data dimensions. Correlation analysis of a pair of indices requires training two models. The process of training the first model is called forward test (FT), which aims to generate realistic index y based on the input index x; the process of training the second model is called reverse test (RT), which aims to generate realistic index x based on the input index y.

For example, in DMs, during FT, residual convolution layers should be adjusted to Conv1d (as y is one-dimensional), and the U-Net architecture should use Conv1d accordingly. While in RT, residual convolution layers are adjusted to Conv2d (as x is two-dimensional), and the U-Net architecture should use Conv2d accordingly.

(d) **Perform FT and RT sequentially:** During model training and validation, use appropriate performance metrics to evaluate the model's effectiveness in FT and RT. For 2D data generation, metrics such as Frechet inception distance (FID), learned perceptual image patch similarity (LPIPS), inception score (IS), and structural similarity index measure (SSIM) can be used. For 1D data generation, metrics like root-mean-square error (RMSE) or mean absolute error (MAE) are commonly used.

(e) **Compare the performance metrics of FT and RT:** Analyze the results to determine the relationship between x and y.

(f) **If performance metrics of one test (FT or RT) consistently outperforms the other:** It suggests that

there is an information dependency between x and y. For example, if FT metrics are better than RT metrics, this implies that x can be transformed into y, but y cannot be reliably transformed back into x, indicating that information of y is also contained within x ($y \in x$).

In this case, x can be "purified" to reduce its information content and explore the inherent correlation of the purified x (x') with y. Now, x is RGB image (H × W × C, where C = 3 and values range from 0 to 255), so it can be purified into grayscale image x' (H × W × C, where C = 1 and values range from 0 to 255). Then, the process from step (c) to step (e) could be implemented for the second time. After the new training round, if FT metrics still outperform RT metrics, grayscale image x' could be purified into binary image x'' (H × W × C, where C = 1 and values are 0 or 1) to implement step (c) to step (e) for the third time. This process is iterated until FT and RT metrics both perform well.

The "purification," i.e., dimensionality reduction process, is aimed at uncovering the fundamental reasons behind correlations between two variables. For instance, when a high-dimensional index (e.g., a RGB image) can serve as a hint to generate a certain low-dimensional index (e.g., numerical values), it implies that the high-dimensional index contains information capable of characterizing that low-dimensional index. While, if the low-dimensional index cannot generate the original high-dimensional data (e.g., the RGB image) conversely, it suggests that the low-dimensional index carries less information than its high-dimensional counterpart. In such cases, applying dimensionality reduction (e.g., converting a RGB image to grayscale or binary image) helps reduce information redundancy in high-dimensional variables. Subsequently, a new round of mutual conversion experiments between the information-reduced high-dimensional indicators and low-dimensional indicators after redundancy elimination. This simplification facilitates a deeper analysis, enabling researchers to systematically dissect the underlying physical relationships between variables across different dimensions, leading to isolating the high-dimensional feature subsets that exhibit strong correlations with low-dimensional index.

(g) **If both FT and RT metrics perform poorly:** This indicates that the two indices cannot be reliably transformed into one another, implying weak inherent correlation between x and y ($x \nsim y$).

(h) **If both FT and RT metrics perform well:** This indicates that the two indices can be reliably transformed into one another, suggesting strong inherent correlation between x and y ($x \sim y$), i.e., information of x and y has significant intersections ($x \cap y$) or effective information density of x and y are equivalent ($x \approx y$).

The following takes DMs as the chosen generative models to explore the inherent correlation of soil properties based on soil image, cone index curve, and TDR waveform to demonstrate the experimental results of the above method.

3 Excavated soil multi-source heterogeneous database

3.1 Excavated soil information collecting system (ESICS)

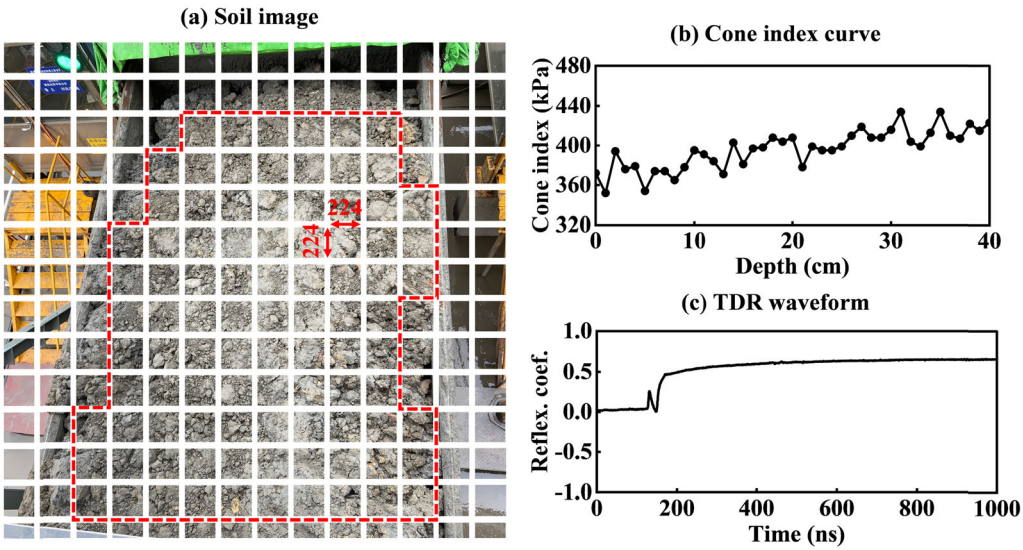
In Xiecun Wharf, the largest platform for transferring excavated soils in China, an excavated soil information collecting system (ESICS) was developed to sample soil multi-source heterogeneous data [21, 56]. The Wharf serves as Hangzhou's centralized trading hub for excavated soils, receiving materials transported by vehicles from nearly 20 foundation and subway projects within a 500-square-kilometer radius. Here, soils are unloaded from trucks and transferred onto vessels for further transportation. By consolidating multi-source soil data at Xiecun Wharf—rather than collecting information separately from scattered construction sites—data can be more diverse and be collected more efficiently. This centralized approach enables a comprehensive analysis of soil types and characteristics across Hangzhou, leading to a representative dataset. Figure 4 shows the configuration and core elements of the ESICS. Installed at the entrance of the Xiecun Wharf, ESICS employed a multi-sensor approach to efficiently gather multi-source heterogeneous data about the excavated soils carried by vehicles. All sensors were assembled into the rapid detection and classification station. Digital cameras were used to capture soil images inside vehicle cars from the overhead views. Time domain reflectometry (TDR) cone penetrometer was used to measure mechanical properties through the cone index curve from the soil surface to 40 cm depth subsoil [55]. Additionally, the TDR cone penetrometer recorded TDR waveforms at 40 cm depth subsoil, which provided insights into moisture content, fine particle content, and even ion compound presence, factors that characterize both engineering and environmental properties of soils. All collected data, including imaging data, mechanical data, and electrical data were displayed on a soil info interactive panel and stored for further analysis. This advanced in situ-tool-based approach allows for accurate identification of soil types and assessment of their quality, which is vital for determining their sustainability for various utilizations.

Fig. 4 Excavated soil information collecting system (ESICS)

The used TDR cone penetrometer consists of four semicircularly shaped stainless-steel conductors which were placed around a 30-mm-diameter poly-ether-ether-ketone (PEEK) shaft through steel nails. The conductors had a diameter of 8 mm and a length of 150 mm. One pair of two opposite conductors was connected to the inner conductor of a 50 Ω-impedance coaxial cable through soldering. The shield of the coaxial cable was connected to the other pair of opposite conductors. Considering the existence of PEEK shaft, the dielectric permittivity and electrical conductivity measured by TDR penetrometer need to be converted into that of soil around the probe through calibration. Mixtures of ethanol (dielectric permittivity is about 16)—deionized water (dielectric permittivity is about 80) with different concentrations (from 0 to 100%, 20% by interval) were selected as target mediums to calibrate the dielectric permittivity. The dielectric permittivity of target medium measured by three-rod TDR probe and the Δt measured by TDR penetrometer can be fitted. The fitted linear relationship between Δt^2 and dielectric permittivity (ε) are fitted as $\varepsilon = -2.878 + 1.50 \cdot \Delta t^2$. The value of R^2 is about 0.99. $CuSO_4$ solution with different concentrations (from 0 to 0.030 mol/L, 0.005 mol/L by interval) was used as target mediums to calibrate the electrical conductivity. The electrical conductivity of target medium measured by three-rod TDR probe and the V_0/V_∞ measured by TDR penetrometer can be fitted. The fitted linear relationship between V_0/V_∞ and electrical conductivity (EC) is fitted as $EC = -125.1 + 206.8 \cdot V_0/V_\infty$. The value of R^2 is about 0.99. A calibration procedure was also developed to correlate the TDR cone penetrometer's force sensor outputs with standard cone penetrometer measurements, deriving an empirical transfer function for cone index estimation.

3.2 Database configuration

The data collection for this study began on October 1, 2021, and lasted for approximately two months. Sampling is conducted daily from 8:00 AM to 4:00 PM. Each day, one channel is randomly selected for sampling, and every soil-transport truck passing through that channel undergoes systematic sampling and data collection. This sampling strategy ensures data diversity and randomness, enabling an accurate representation of the distribution patterns of excavated soil types. A total of 3243 data groups were collected, with each group consisting of one soil original image (Fig. 5a), one cone index (CI) curve (Fig. 5b), and one TDR waveform (Fig. 5c). The original soil image (Fig. 5a) captures the entire soil inside a vehicle. To enhance the dataset and better represent small-scale features of soil surface morphology, a data augmentation strategy was applied. In detail, the soil original image in Fig. 5a was segmented into 93 sub-images of size 224 × 224 like Fig. 5a. Due to the fact that all these 93 soil sub-images are from the same vehicle, they also share the same cone index curve, TDR waveform, and soil information record. In this way, one piece of data can be



Index type	Data type	Dimension	Details
Soil image	Imaging data	2-D	224×224 pixel RGB image
Cone index	Spatial-series data	1-D	Single-column array containing 41 elements
TDR waveform	Time-series data	1-D	Single-column array containing 2,048 elements

Fig. 5 Database configuration

expanded into 93 pieces of data, helping the model to learn the morphological features of soil accurately. This data augmentation strategy expands 3243 groups of original data into 23,122 groups, enabling the model to better learn soil morphological features.

The CI curve represents cone index values across depths from 0 to 40 cm in the form of single-column array containing 41 elements (Fig. 5b). This data exhibits strong spatial-series characteristics, as adjacent values are physically related. Therefore, CI curves are suggested to be analyzed holistically rather than as isolated data points. Similarly, the TDR waveform is time-series data consisting of 2048 data points over approximately 1000 ns (Fig. 5c). This data is frequently used to estimate soil moisture content and electrical conductivity, which are linked to soil texture, fine particle content, and ion compound content. Like the CI curve, the TDR waveform also requires holistic analysis to preserve its temporal correlations.

The ESICS at Xiecun Wharf exemplifies a sophisticated data collection system that enhances the efficiency of soil information gathering. By developing the multi-source heterogeneous database using advanced techniques, it not only facilitates the processing of vast quantities of geomaterials but also contributes to a deeper understanding of soil characteristics.

The particle size distribution (PSD) of soil samples is shown in the soil texture triangle of Fig. 6. Based on Ayers et al. [5], the USCS classification can map onto the United States Department of Agriculture (USDA) triangle and

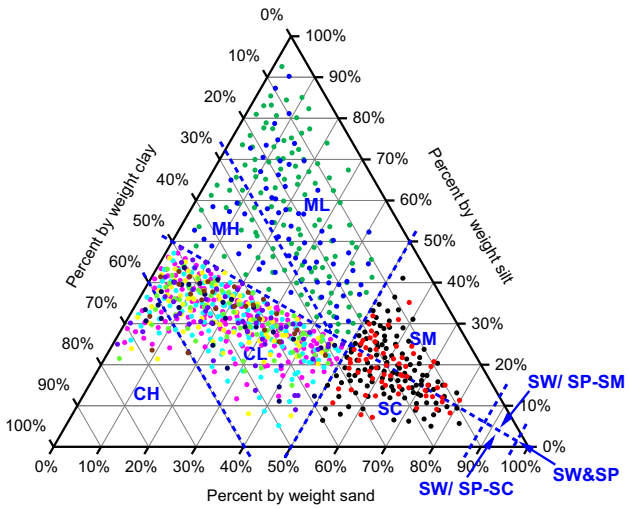


Fig. 6 Soil types in the USDA&USCS soil triangle

divide soils into nine types as clean sands (SW and SP), clean sands to sands with fines of silts (SW/SP-SM), clean sands to sands with fines of clay (SW/SP-SC), sands with fines of silts (SM), sands with fines of clay (SC), lean silt (ML), elastic silt (MH), lean clay (CL), and fat clay (CH). Obviously, the soil samples in this database were dominated by fine-grained soils, thus the generalization ability of the model in fine-grained soils can be guaranteed.

4 Case studies

4.1 Case design

To study the inherent correlation between soil images (visual information), cone index curves (mechanical indicators) [31], and TDR waveforms (electrical indicators), three cases were designed to investigate the modal transformation of these multi-source heterogeneous data, as shown in Fig. 7:

Case 1 aims to investigate the inherent correlations between soil images and cone index curves. The soil RGB color image was used as the input data to generate the cone index curve (forward test); the cone index curve was then used as the input data to generate the soil RGB color image (first-round reverse test); next, the RGB images in the dataset were purified into grayscale images. The cone index curve was used as input data to generate the soil grayscale images (second-round reverse test); finally, the grayscale images were further purified into binary images. The cone index curve was used as input data to generate the soil binary images (third-round reverse test).

Case 2 aims to investigate the inherent correlations between soil images and TDR waveforms. The soil RGB color image was used as the input data to generate the TDR waveform (forward test); the TDR waveform was then used as the input data to generate the soil RGB color image (first-round reverse test); next, the RGB images in the dataset were purified into grayscale images. The TDR

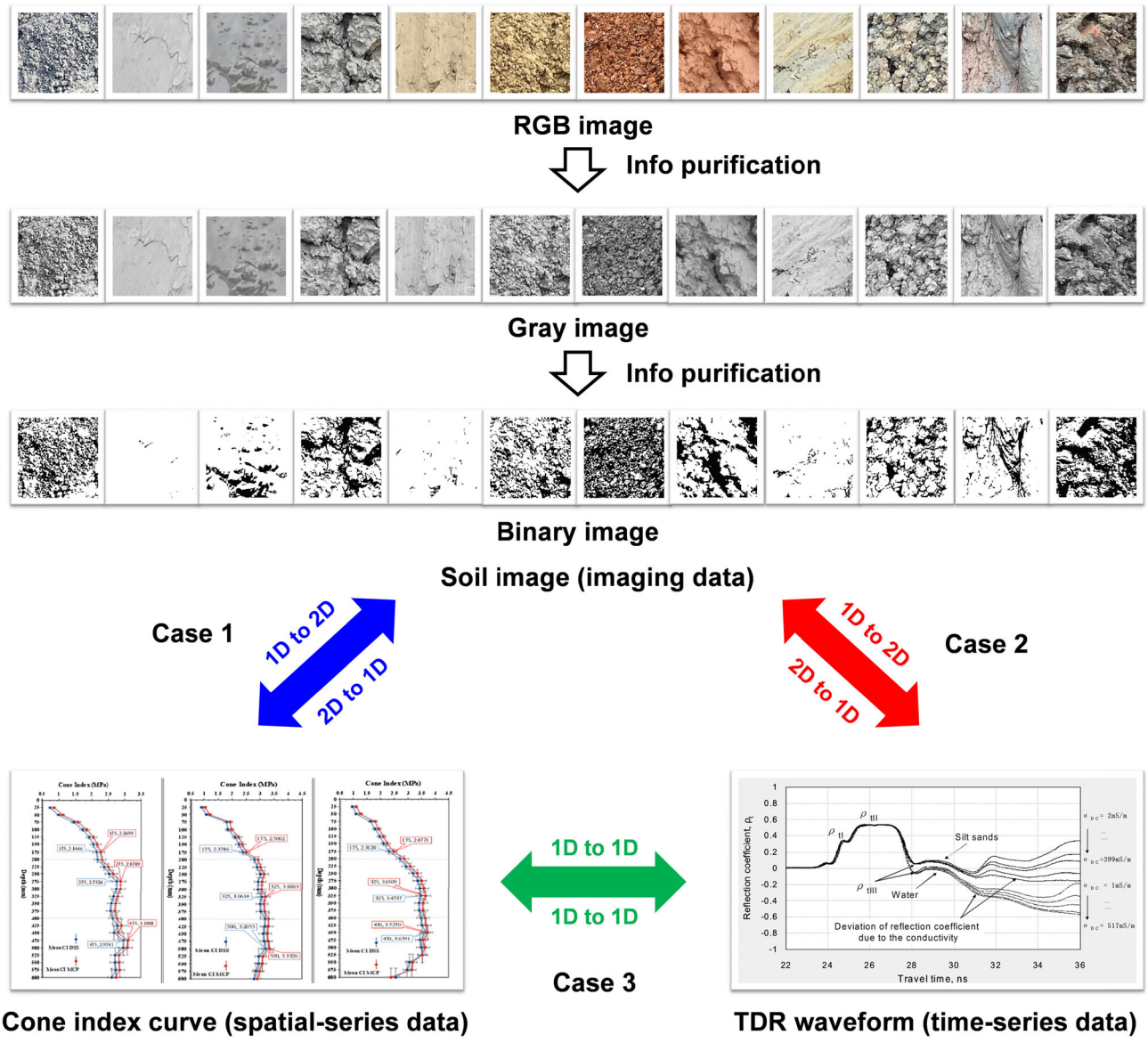


Fig. 7 Case of multimodal transition

waveform was used as input data to generate the soil grayscale images (second-round reverse test); finally, the grayscale images were further purified into binary images. The TDR waveform was used as input data to generate the soil binary images (third-round reverse test).

Case 3 aims to investigate the inherent correlations between cone index curves and TDR waveforms. The cone index curve was used as the input data to generate the TDR waveform (forward test); the TDR waveform was then used as the input data to generate the cone index curve (reverse test).

The generative-model-based experiments involved the transformation between 2D data (soil images) and 1D data (cone index curves and TDR waveforms), as well as the transformation between 1D datasets. These tests ultimately revealed the inherent relationships among the three types of soil indices in different modalities.

4.2 DDPM architecture

Figure 8 illustrates the framework of the denoising diffusion probabilistic models (DDPMs) used in the experiment [6, 40, 42]. In the forward process, noise is incrementally added to the data until the original data becomes Gaussian distribution. At each time step t, the data point was sampled from the Gaussian distribution $q(x_t|x_{t-1})$ derived from the previous time step x_{t-1}.

$$q(x_t|x_{t-1}) := N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

where $\beta_t \in (0,1)$ is the variance schedule; I is the identity matrix.

In the reverse process, the goal is to denoise images iteratively to get an image with less noise. A time-dependent function approximator was employed to predict the Gaussian distribution $p(x_{t-1}|x_t)$ at each step.

$$p(x_{t-1}|x_t) := N(x_{t-1}; \widetilde{\mu}_t(x_t,t), \sigma_t I) \tag{2}$$

$$\widetilde{\mu}_t(x_t,t) \approx \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_t\right) \tag{3}$$

$$\alpha_t := 1 - \beta_t \tag{4}$$

where ε_t is the noise introduced in step t. The model aims to learn this distribution to denoise images by reverse conditional probability. So, a neural network $\varepsilon_\theta(x_t, t)$ should be trained to approximate the introduced noise distribution. The model architecture for this process is typically a U-Net. During the sampling phase, sinusoidal encoding was used to encode the timestep t, and suitable embedders were utilized to pass prompts into the model. For example, in the task of generating soil images from cone index curves, the cone index curve was encoded as a factor using an LSTM for cone index curves representing spatial-series data. Similarly, in the task of generating soil images from TDR waveforms, LSTM was also used to encode the time-series data of the TDR waveform. If the input data is an image, the corresponding encoder should be CNN. The U-Net structure consisted of down-sampling and up-sampling streams, connected by skip connections to merge shallow and deep features. In the encoding part (down-sampling), blocks of max pooling are followed by convolutional

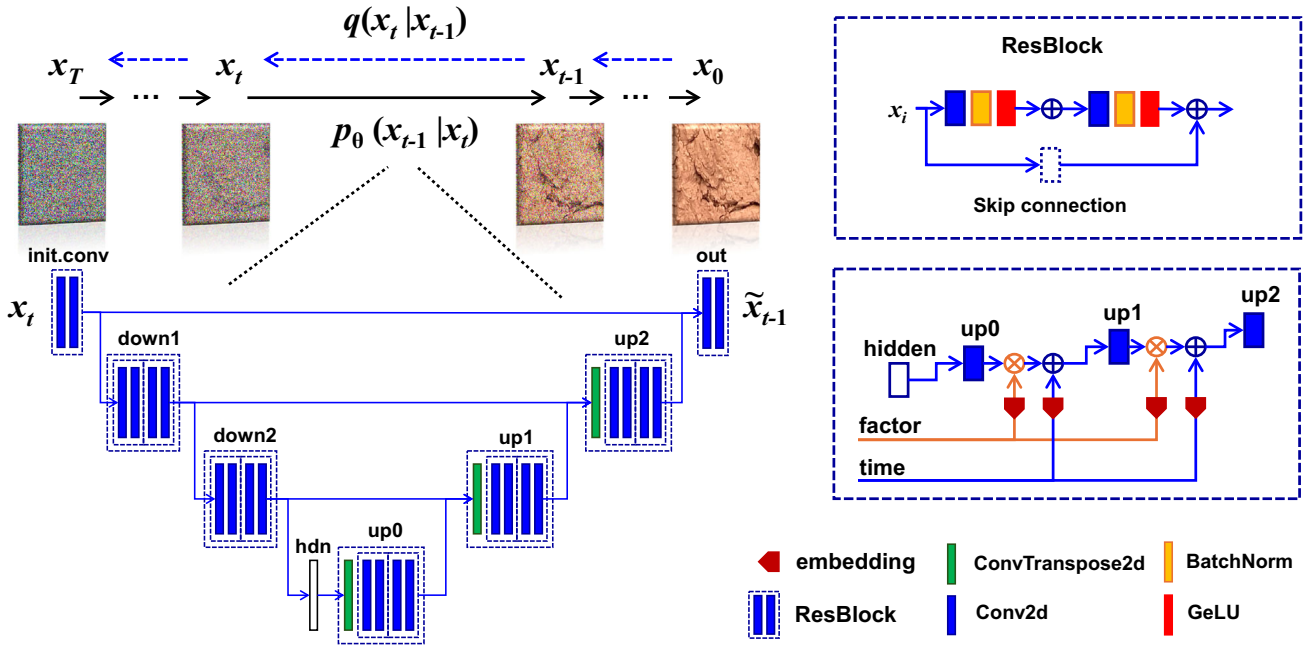


Fig. 8 DDPM architecture

layers, group normalization, and GELU (Gaussian error linear unit) activation functions. The decoding part (up-sampling) mirrors this structure, using bilinear up-sampling followed by convolutional layers, group normalization, and GELU activation functions. To enhance training stability and gradient flow, ResNet blocks are incorporated into the sampling modules. Each layer in the framework includes two residual blocks, with embeddings passed into each block. Additionally, down-sampled, up-sampled, and pre-residual values are returned and stored for use in residual concatenated skip connections. This design ensures efficient feature propagation, enabling the model to effectively capture hierarchical information and represent complex data distributions. Specific algorithm parameters are shown in Tables 1 and 2, as well as some hyperparameters are shown in Table 3.

5 Model training settings and performance metrics

During the model training, the number of sampling time-steps was set to 400. To reduce unnecessary computational costs, the image resolution in the dataset was resized from 224 × 224 to 84 × 84. The number of epochs was set to 100, with an initial learning rate of 1e−3. A linearly decaying learning rate was used to progressively adjust the learning rate, and the Adam optimizer was employed. The MSE loss was selected as the loss function.

To evaluate the effectiveness of the trained model, it is essential to select appropriate performance metrics. For image generation tasks, the metrics used were structural similarity index measure (SSIM) and learned perceptual image patch similarity (LPIPS).

SSIM measures the structural similarity between the generated and ground truth images, reflecting the preservation of spatial details and perceptual quality.

Table 1 Algorithm parameters in case generating images based on cone index

Component	Layer type	Parameters	Input shape	Output shape	Activation
Residual ConvBlock	Conv2d + BN + GELU	in_chan, out_chan, kernel = 3 × 3, stride = 1, padding = 1	B, in_c, H, W	B, out_c, H, W	GELU
	Conv2d + BN + GELU	out_chan, out_chan, kernel = 3 × 3, stride = 1, padding = 1	B, out_c, H, W	B, out_c, H, W	GELU
UnetDown	Residual ConvBlock × 2	in_chan → out_chan	B, in_c, H, W	B, out_c, H/2, W/2	–
	MaxPool2d	kernel = 2, stride = 2	B, out_c, H, W	B, out_c, H/2, W/2	/
UnetUp	Conv Transpose2d	in_chan, out_chan, kernel = 2 × 2, stride = 2	B, in_c, H, W	B, out_c, 2H, 2W	–
	Residual ConvBlock × 2	out_chan → out_chan	B, out_c, 2H, 2W	B, out_c, 2H, 2W	GELU
Embed LSTM	LSTM	input_size = 2, hidden_size = emb_dim, num_layers = 2, batch_first = True	B, seq_len = 40, 2	B, emb_dim	GELU
EmbedFC	Linear + GELU + Linear	input_dim → emb_dim → emb_dim	B, input_dim	B, emb_dim	GELU
CiUnet	Initial Conv	in_channels = 3 → n_feat = 64	B, 3, 84, 84	B, 64, 84, 84	–
	Down1	64 → 64	B, 64, 84, 84	B, 64, 42, 42	–
	Down2	64 → 128	B, 64, 42, 42	B, 128, 21, 21	–
	AvgPool2d	kernel = 4	B, 128, 21, 21	B, 128, 5, 5	GELU
	Up0 (Transpose)	128 → 128, kernel = 5, stride = 4	B, 128, 5, 5	B, 128, 21, 21	RELU
	Up1	256 → 64 (with skip from Down2)	B, 256, 21, 21	B, 64, 42, 42	–
	Up2	128 → 64 (with skip from Down1)	B, 128, 42, 42	B, 64, 84, 84	–
	Output Conv	128 → 64 → 3	B, 128, 84, 84	B, 3, 84, 84	–

Table 2 Algorithm parameters in case generating cone index based on images

Component	Layer type	Parameters	Input shape	Output shape	Activation
Residual ConvBlock	Conv1d + GELU	in_chan, out_chan, kernel = 3, stride = 1, padding = 1	B, in_c, L	B, out_c, L	GELU
	Conv1d + GELU	out_chan, out_chan, kernel = 3, stride = 1, padding = 1	B, out_c, L	B, out_c, L	GELU
UnetDown	Residual ConvBlock × 2	in_chan → out_chan	B, in_c, L	B, out_c, L// 2	–
	MaxPool1d	kernel = 2, stride = 2	B, out_c, L	B, out_c, L// 2	–
UnetUp	Conv Transpose1d	in_chan, out_chan, kernel = 2, stride = 2	B, in_c, L	B, out_c, 2L	–
	Residual ConvBlock × 2	out_chan → out_chan	B, out_c, 2L	B, out_c, 2L	GELU
EmbedCNN	ResNet18 + Linear	Pretrained ResNet18 → Linear (512 → emb_dim)	B, 3, 224, 224	B, emb_dim	RELU
EmbedFC	Linear + GELU + Linear	input_dim → emb_dim → emb_dim	B, input_dim	B, emb_dim	GELU
WaveUnet	Initial Conv	in_channels = 1 → n_feat = 64	B, 1, 40	B, 64, 40	–
	Down1	64 → 64	B, 64, 40	B, 64, 20	–
	Down2	64 → 128	B, 64, 20	B, 128, 10	–
	AvgPool1d	kernel = 4	B, 128, 10	B, 128, 5	GELU
	Up0Transpose	128 → 128, kernel = 2, stride = 2	B, 128, 5	B, 128, 10	RELU
	Up1	256 → 64 (with skip from Down2)	B, 256, 10	B, 64, 20	–
	Up2	128 → 64 (with skip from Down1)	B, 128, 20	B, 64, 40	–
	Output Conv	128 → 64 → 1	B, 128, 40	B, 1, 40	–

Table 3 Training hyperparameter settings

Parameter	CI → img		img → CI	
	Value	Description	Value	Description
timesteps	600	Diffusion steps	400	Diffusion steps
n_feat	64	Base channel dimension	64	Base channel dimension
n_cifeat	10	CI feature dimension	100	Image feature dimension
height	84	Input image height/width	40	Input waveform length
batch_size	32	Training batch size	32	Training batch size
lrate	1e−3	Initial learning rate	1e−3	Initial learning rate
beta1/beta2	1e−4/ 0.02	Noise schedule bounds	1e−4/ 0.02	Noise schedule bounds

$$\mathrm{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)} \tag{5}$$

where μ_x and μ_y are the mean values of image x and image y; σ_x and σ_y are the variances of image x and image y; σ_{xy} is the covariance between the images; C_1 and C_2 are constants to prevent division by zero. The SSIM value ranges between − 1 and 1. When SSIM = 1, it indicates that the two images are identical in terms of structure, luminance, and contrast. When SSIM < 0, it suggests that the generated image has poor quality and a significant deviation from the reference image. In general, higher SSIM values indicate better structural similarity and perceptual quality between the compared images.

LPIPS measures the perceptual similarity between two images based on deep network features. Unlike SSIM, LPIPS focuses on the similarity of images in the deep feature space, which better aligns with human visual perception. By using a pre-trained CNN (such as VGG or AlexNet), the deep features of the images are extracted, and the distance between these features is calculated:

$$\mathrm{LPIPS}(x,y)=\sum_l \frac{1}{H_l W_l}\sum_{h,w}\left\|f_l^x(h,w)-f_l^y(h,w)\right\|_2^2 \tag{6}$$

where f_l^x and f_l^y are the features of the l-th layer in the network for image x and image y; H_l and W_l are the height and width of the feature map at the l-th layer; $\|a-b\|_2$ denotes the Euclidean distance. The LPIPS typically falls in the range of 0 to 1. When LPIPS = 0, it indicates that the

two images are identical in terms of perceptual similarity (minimum perceptual distance). When LPIPS $\approx$ 1, it indicates that the two images are very different perceptually (maximum perceptual distance). A lower LPIPS value indicates that the two images are more similar and perceptually closer.

For waveform generation tasks, the performance metrics adopted were root-mean-square error (RMSE) and mean absolute error (MAE).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

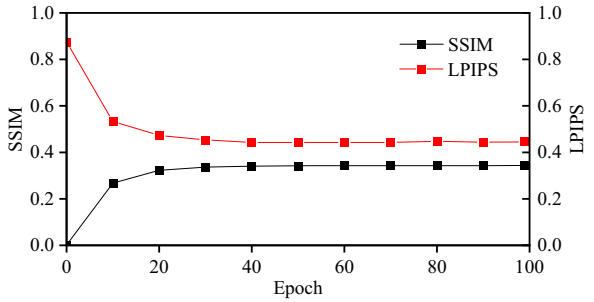
where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the sample amount.

The selection of SSIM, LPIPS, RMSE, and MAE as evaluation metrics provides a comprehensive framework for assessing generated data quality from multiple perspectives. For image data, SSIM is used to measure structural similarity between generated and real data (luminance, contrast, structure), while LPIPS quantifies perceptual similarity using deep features (VGG/AlexNet) and captures high-level semantic differences (e.g., texture patterns, morphological features). SSIM and LPIPS work synergistically to diagnose data errors in generated data from the perspectives of local structural distortion, global deviation, and semantic anomalies. Moreover, SSIM is based on a physiological model of the human visual system (HVS), and LPIPS encodes object recognition prior knowledge acquired from ImageNet training, thus providing physical interpretability. For waveform data, RMSE and MAE are used to provide quantitative rigor for deterministic tasks and validate whether generated data meet physical constraints, enabling error diagnosis from the perspectives of global deviation and numerical drift. When the generated index y is a two-dimensional, performance metrics such as SSIM and LPIPS can be chosen to assess the degree of correlation; when y is one-dimensional, the RMSE and MAE can be chosen for correlation assessment.

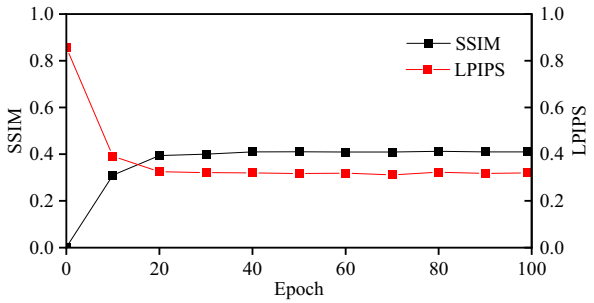
6 Results and analysis

6.1 Image and cone index curve

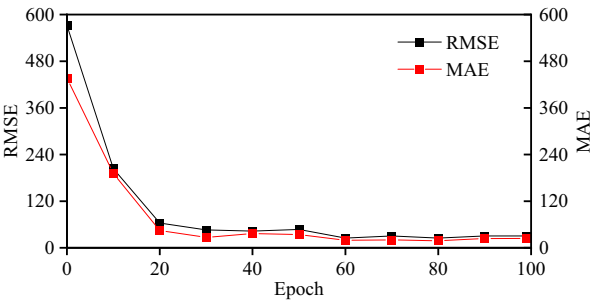
This case was used to explore the inherent correlation between soil images and cone index curves. Figure 9 shows the performance metrics (SSIM is black and LPIPS is red) during the training of the FT and RT models. In the experiment of generating RGB soil images based on cone



(a) Case of colorful soil image generation based on cone index curve



(b) Case of binary soil image generation based on cone index curve



(c) Case of cone index curve generation based on colorful soil image

Fig. 9 Performance metrics of model training (img ↔ CI)

index curves (Fig. 9a), the training performance stabilized after approximately 20 epochs. The final SSIM was around 0.34 and LPIPS was around 0.45, indicating that the pixel similarity between the soil RGB images generated based on cone index curves and the real images was low, and there were significant perceptual differences. The overall image quality was not good. In the experiment of generating binary soil images based on cone index curves (Fig. 9b), the training performance also stabilized after approximately 20 epochs. The final SSIM was around 0.41 and LPIPS was around 0.32, suggesting that after removing color-related information from the soil images, the model's training performance improved. The pixel similarity between the binary soil images generated based on cone index curves and the real binary images was high, and the perceptual differences were significantly reduced. In the experiment of generating cone index values based on soil

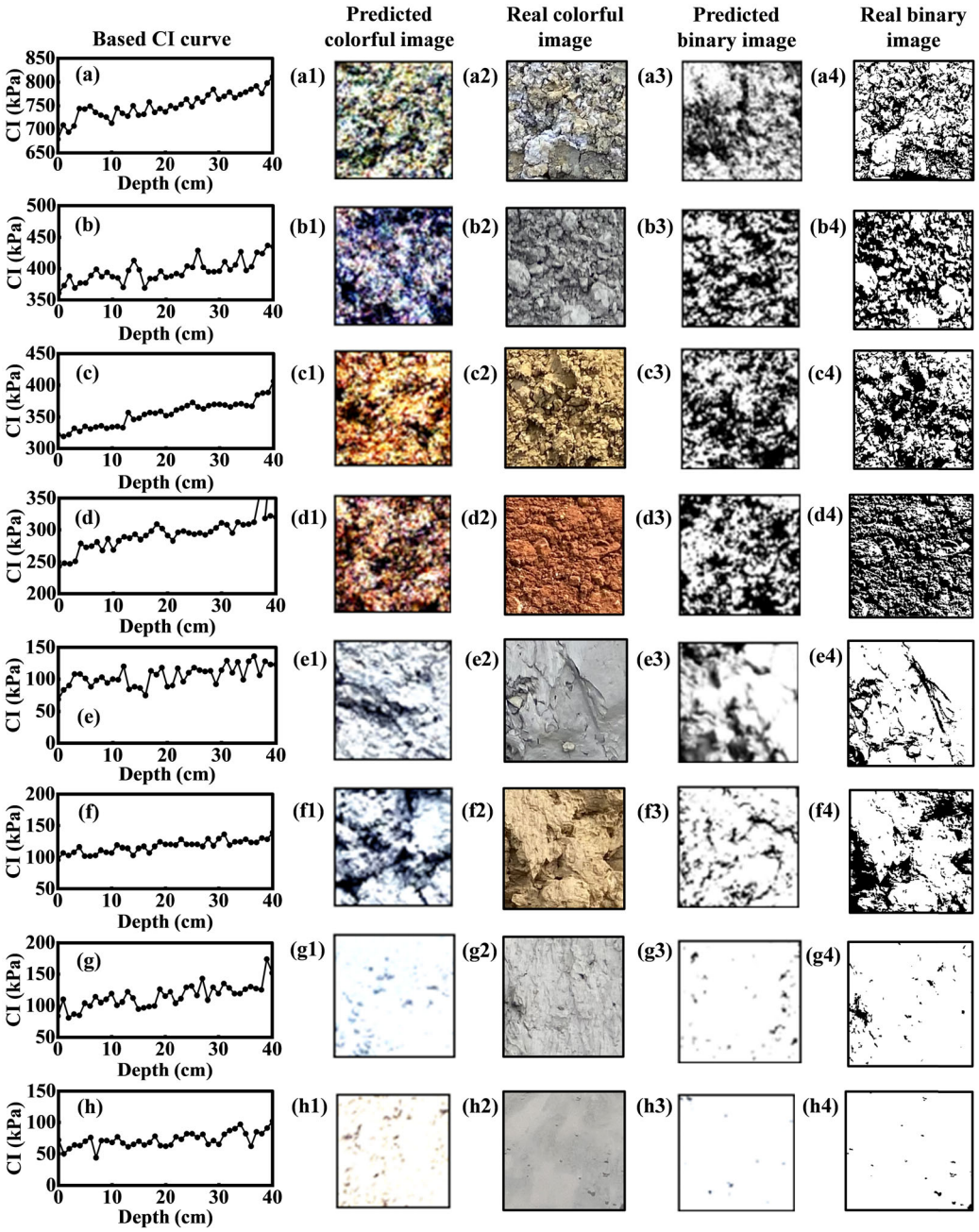


Fig. 10 Generated colorful soil images based on cone index curve

RGB images (Fig. 9c), the training performance stabilized after approximately 20 epochs. The final RMSE and MAE reached 30.6 and 24.1, respectively, indicating that the average deviation between the predicted cone index values and the actual cone index values was around 30.

Figure 10 shows the model testing results for generating soil images based on cone index curves. From left to right, there are based CI curves, the predicted colorful images, the real colorful images, the predicted binary images, and the real binary images. When predicting RGB images, the model exhibited significant uncertainty. For coarse-grained soils with CI values above 200 kPa, Fig. 10a1, Fig. 10c1, and Fig. 10d1 were perceptually similar to their corresponding real images Fig. 10a2, Fig. 10c2, and Fig. 10d2. However, there was a significant color difference between Fig. 10b1 and its corresponding ground truth Fig. 10b2. For fine-grained soils with CI values below 200 kPa, Fig. 10e1 and Fig. 10g1 were perceptually similar to their corresponding real images Fig. 10e2 and Fig. 10g2, but Fig. 10f1 and Fig. 10h1 show considerable color

differences compared to the real images Fig. 10f2 and Fig. 10h2. Obviously, it is uncertain to generate RGB images based on cone index curves, indicating that the information contained in cone index curves does not include soil color characteristics. Therefore, when soil color information was removed, the similarity of binary images generated from cone index curves improved significantly. Figure 10a3–h3 generally reflects the apparent morphologies of their corresponding real soil images Fig. 10a4–h4. There is a noticeable discrepancy between the RGB images generated using the cone index curve as a hint and the actual RGB images. This implies that the cone index does not capture information about the soil's mineral composition, making it difficult to accurately represent the soil's color characteristics. Consequently, the predicted colorful images exhibit disordered color information. However, the cone index data can effectively reflect textural features such as particle size and moisture content. As a result, the predicted binary images generated based on the cone index show a high degree of similarity to the real binary images. This is because binary images discard color information, allowing the cone index's inherent ability to characterize soil properties to be more effectively excited.

Figure 11 illustrates the model testing results for generating cone index curves based on soil RGB images. On the left are the based colorful images, and on the right are the predicted CI curves (black) compared with the real CI curves (red). To ensure the physical significance of the generated cone index curves, normalization was applied in the training code to guarantee all cone index values remain positive. It can be observed that the real cone index curves generally exhibited a trend of gradual increase with depth. The predicted cone index curves roughly replicated this trend and fluctuated within the range of the real curves. However, the predicted curves showed significant data volatility. For example, in Fig. 11a, the range of the predicted curve was 168.5 kPa, while the real curve's range was 100 kPa. In Fig. 11b, the ranges were 181.4 kPa and 80 kPa. In Fig. 11c, the ranges were 199 kPa and 78 kPa. In Fig. 11d, the ranges were 161.3 kPa and 76 kPa. In Fig. 11e, the ranges were 131.4 kPa and 66 kPa. In Fig. 11f, the ranges were 87.1 kPa and 54 kPa. Moreover, the predicted curve's maximum and minimum values appeared more randomly compared to the real curves, where the maximum values typically occurred at the deepest depth and the minimum values at the shallowest depth. It can be inferred that soil colorful images contain mechanical characteristics related to the cone index (as the range and trends of the predicted curves were generally consistent with the real curves). However, impurities and random pixels in the images can affect the accuracy of the generated cone index curves, leading to fluctuations and discrepancies in the predicted values.

6.2 Image and TDR waveform

This case was used to explore the inherent correlation between soil images and TDR waveforms. Figure 12 shows the performance metrics during the training of the FT and RT models. In the experiment of generating RGB soil images based on TDR waveforms (Fig. 12a), the training performance stabilized after approximately 40 epochs. The final SSIM was around 0.31 and LPIPS was around 0.42, indicating that the pixel similarity between the soil RGB images generated based on TDR waveforms and the real images was also low, and there were significant perceptual differences. The overall image quality was not good. In the experiment of generating binary soil images based on TDR waveforms (Fig. 12b), the training performance stabilized after approximately 40 epochs. The final SSIM was around 0.40 and LPIPS was around 0.35, suggesting that after removing color-related information from the soil images, the model's training performance also improved. The pixel similarity between the binary soil images generated based on TDR waveforms and the real binary images was high, and the perceptual differences were significantly reduced. In the experiment of generating TDR waveforms based on soil RGB images (Fig. 12c), the training performance stabilized after approximately 20 epochs. The final RMSE and MAE reached 0.21 and 0.17, respectively, indicating that the average deviation between the predicted TDR waveforms and the actual TDR waveforms was around 0.20.

Figure 13 shows the model testing results for generating soil images based on TDR waveforms. From left to right, there are based TDR waveforms, the predicted colorful images, the real colorful images, the predicted binary images, and the real binary images. When predicting RGB images, the model also exhibited significant uncertainty. For samples with lower water content and higher coarse particle content (Fig. 13a1–d1) (samples whose peak and valley in the TDR waveform between 100 and 200 ns have a shorter time interval), the predicted colorful images might exhibit significant color distortion or even generate anomalous samples with abnormal colors, such as Fig. 13b1. In contrast, for samples with higher water content and higher fine particle content (Fig. 13e1–h1) (samples whose peak and valley in the TDR waveform between 100 and 200 ns have a larger time interval), there were significant overexposures on smooth surfaces. Additionally, when the soil's electrical conductivity was excessively high (the ultimate voltage of the TDR waveform was too low), the generated colorful image almost lack visible pixels, as shown in Fig. 13h1. Similar to the patterns observed in images generated based on CI curves, once the RGB images in the training dataset were cleared, the similarity between the predicted binary images and the real

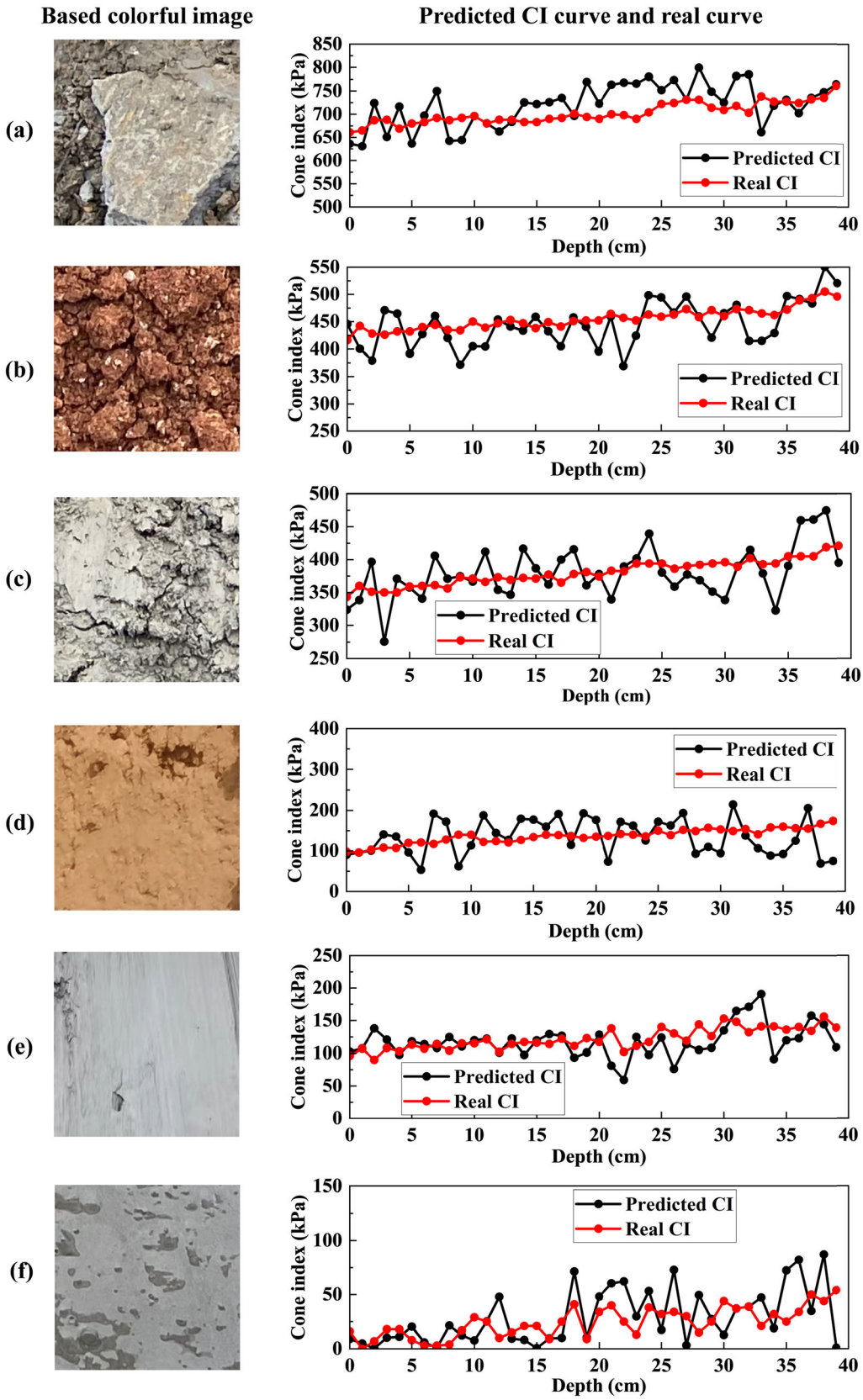
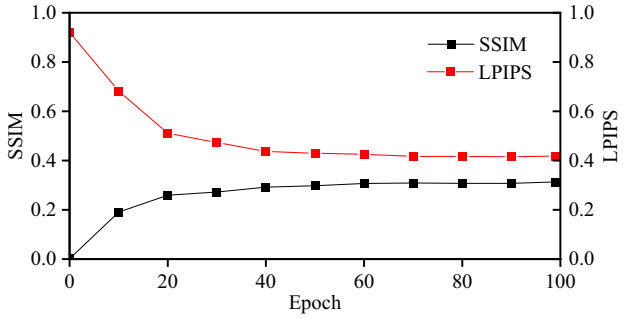
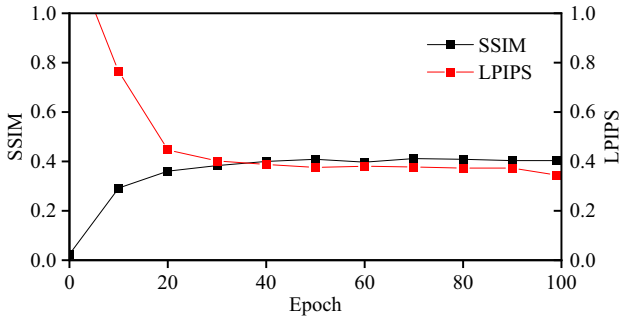


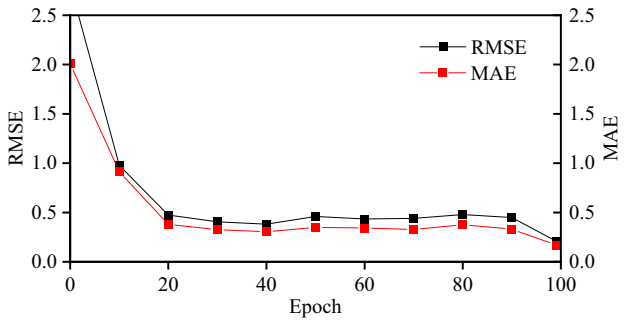
Fig. 11 Generated cone index curve based on colorful soil images



(a) Case of colorful soil image generation based on TDR waveform



(b) Case of binary soil image generation based on TDR waveform



(c) Case of TDR waveform generation based on colorful soil image

Fig. 12 Performance metrics of model training (img ↔ TDR)

images generated based on TDR waveforms improved significantly. The pixel distribution in the predicted binary images had similarity to that in the real soil images.

Figure 14 illustrates the TDR waveforms generated based on colorful images. On the left are the based colorful images, and on the right are the predicted TDR waveforms compared to the real waveforms. Generally, the predicted waveforms appeared relatively smooth, while the real waveforms exhibited slight fluctuations due to impedance changes inside the cable of the detection equipment. For the time interval between peak and valley in the 0–200 ns, which reflects soil water content information (permittivity of soil), the predictions demonstrated relatively good accuracy. Across samples with water content ranging from low (Fig. 14a) to high (Fig. 14e), there were cases where the predictions were consistent with the real waveforms. However, there were significant discrepancies in the prediction of electrical conductivity, represented by the ultimate voltage in the waveform. Some models overestimated ultimate voltages, while some underestimated. For example, in Fig. 14a, the ultimate voltage of the predicted waveform was 0.64, while the ultimate voltage of the real waveform was 0.52. In Fig. 14d, the ultimate voltage of the predicted waveform was 0.12, while the ultimate voltage of the real waveform was -0.12. It can be found that in cases of overestimation, the differences between the estimation and the real values were particularly significant. This is because soil may contain contaminants or ionic compounds, leading to excessively high electrical conductivity. However, such contamination information cannot be characterized in soil images, making it challenging for TDR waveforms generated from images to accurately predict real electrical conductivity.

6.3 Cone index curve and TDR waveform

This case was used to explore the inherent correlation between cone index curves and TDR waveforms. Figure 15 shows the performance metrics during the training of the FT and RT models. In the experiment of generating cone index curves based on TDR waveforms (Fig. 15a), the training performance stabilized after approximately 30 epochs. The final RMSE was around 27.7, and MAE was around 20.9, indicating that the average deviation between the predicted CI curves and the actual CI curves was around 20–30. In the experiment of generating TDR waveforms based on CI curves (Fig. 15b), the training performance stabilized after approximately 30 epochs. The final RMSE was around 0.32, and MAE was around 0.24, suggesting that the average deviation between the predicted TDR waveforms and the actual TDR waveforms was around 0.30. The training performance of the models for converting the above two types of 1D waveforms into each other is roughly equivalent, indicating a clear inherent correlation between them.

Figure 16 shows the model testing results for generating CI curves based on TDR waveforms. On the left are the based TDR waveforms, and on the right are the predicted CI curves compared to the real curves. It can be found that the fluctuation range of the predicted curves was roughly equivalent to that of the real curves and showed the same trend of gradual increase with depth. However, the predicted CI curves exhibited greater volatility and uncertainty compared to the real ones. For example, in Fig. 16a, the range of the predicted curve was 279.4 kPa, while the real curve's range was 72 kPa. In Fig. 16b, the ranges were 219.1 kPa and 78 kPa. In Fig. 16c, the ranges were 157.5 kPa and 108 kPa. In Fig. 16d, the ranges were

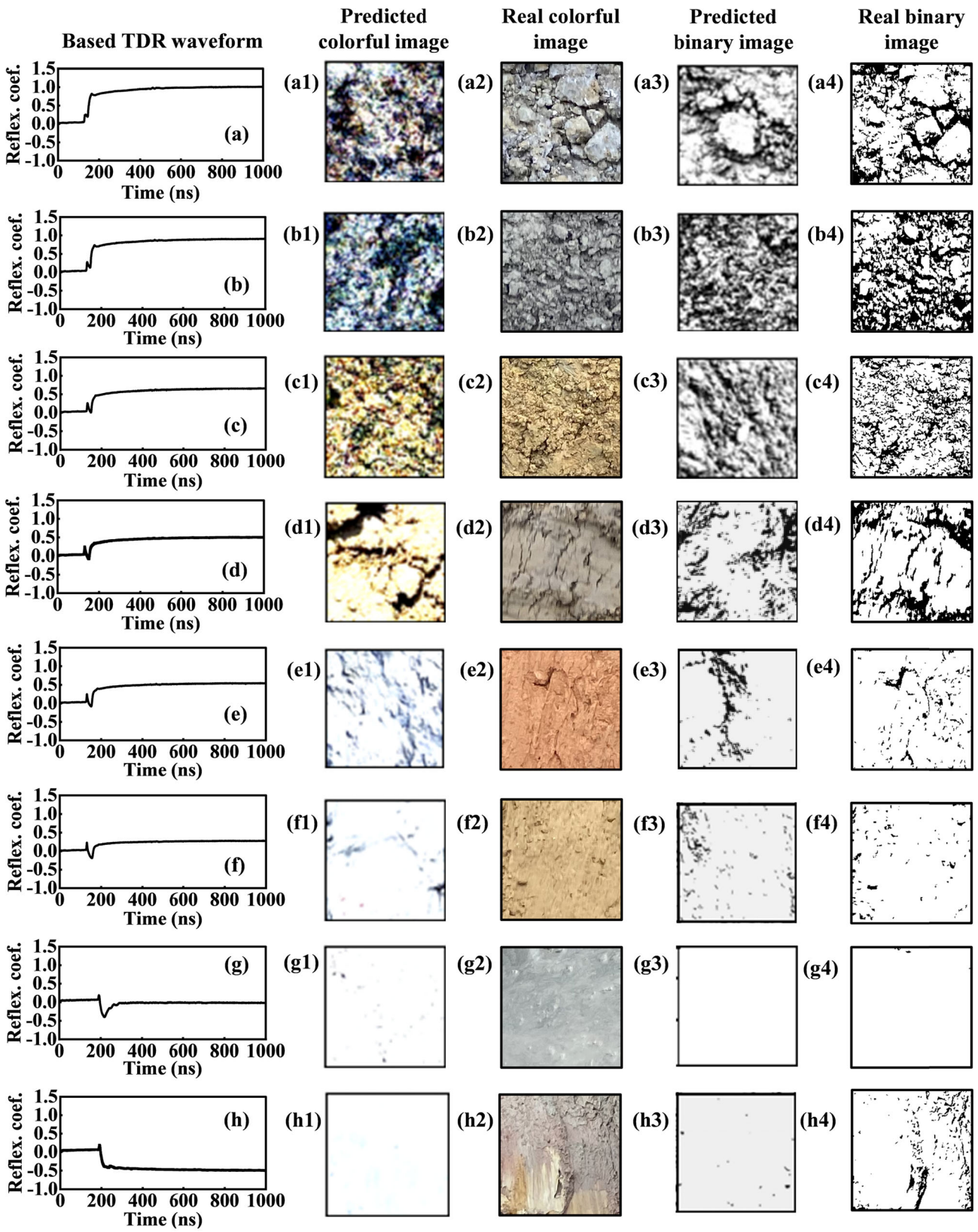


Fig. 13 Generated colorful soil images based on TDR waveform

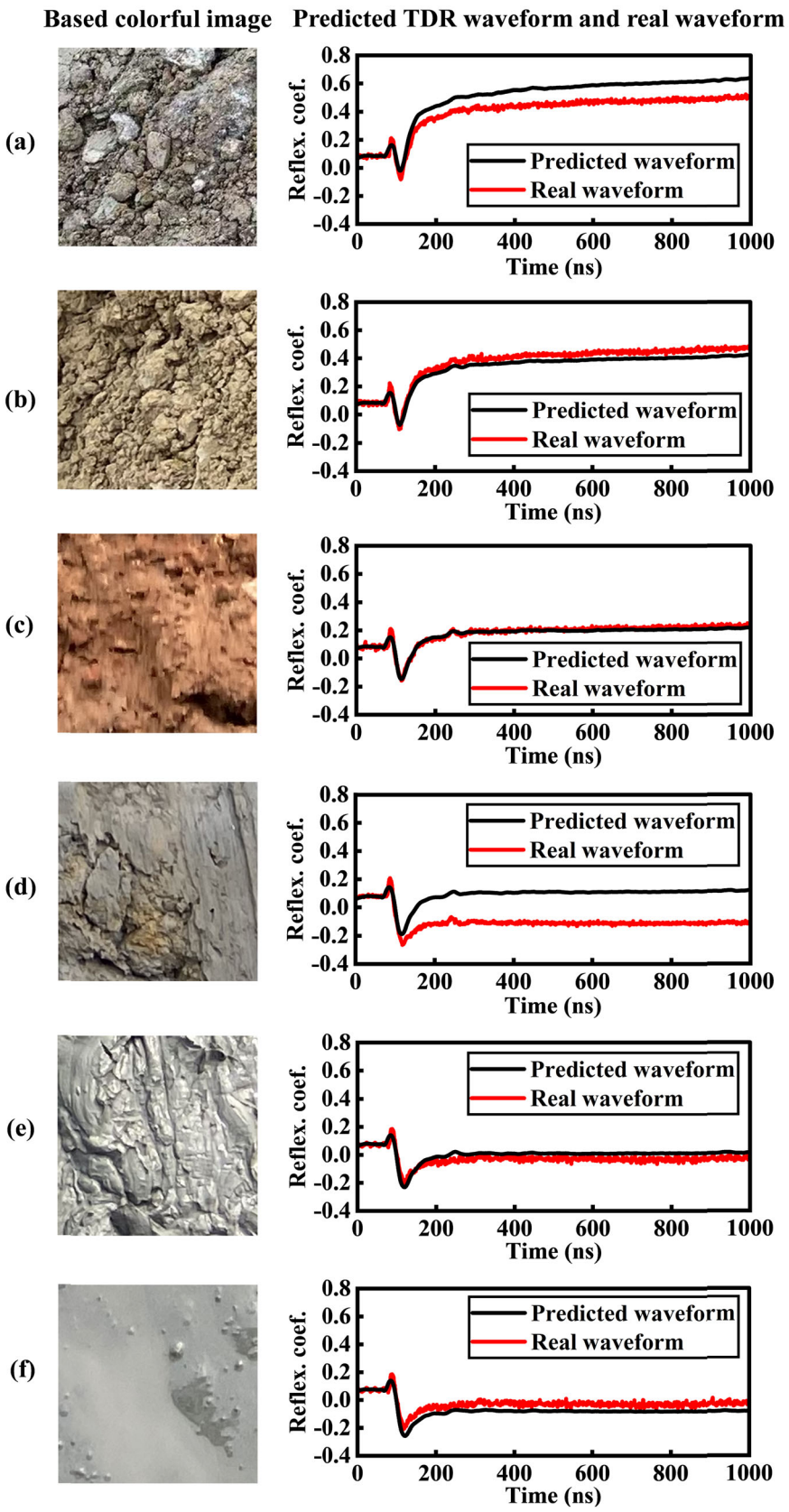
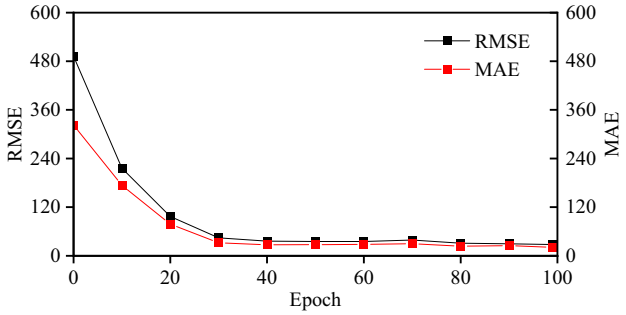
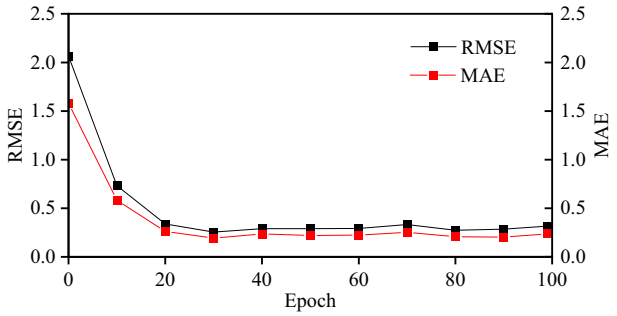


Fig. 14 Generated TDR waveform based on colorful soil images



(a) Case of cone index curve generation based on TDR waveform



(b) Case of TDR waveform generation based on cone index curve

Fig. 15 Performance metrics of model training (CI ↔ TDR)

146.5 kPa and 54 kPa. In Fig. 16e, the ranges were 122.1 kPa and 53 kPa. In Fig. 16f, the ranges were 89.6 kPa and 67 kPa. Moreover, the predicted curve's maximum and minimum values appeared more randomly compared to the real curves, where the maximum values typically occurred at the deepest depth and the minimum values at the shallowest depth. It can be inferred that soil TDR waveforms contain mechanical characteristics related to the cone index (as the range and trends of the predicted curves were generally consistent with the real curves). However, noises in the TDR waveforms and the voltage changes induced by environmental factors (e.g., ion compounds) can affect the accuracy of the generated CI curves, leading to fluctuations and discrepancies in the predicted values.

Figure 17 illustrates the TDR waveforms generated based on CI curves. On the left are the based CI curves, and on the right are the predicted TDR waveforms compared to the real waveforms. Generally, the predicted waveforms appeared relatively smooth, while the real waveforms exhibited slight fluctuations due to impedance changes inside the cable of the detection equipment. For the time interval between peak and valley in the 0–200 ns, which reflects soil water content information (permittivity of soil), the predictions demonstrated relatively good accuracy. Across samples with water content ranging from low (Fig. 17a) to high (Fig. 17d), there were cases where the predictions were consistent with the real waveforms. However, there were also significant discrepancies in the prediction of electrical conductivity, represented by the ultimate voltage in the waveform. Some models overestimated ultimate voltages. For example, in Fig. 17d, the ultimate voltage of the predicted waveform was 0.44, while the ultimate voltage of the real waveform was 0.30. In Fig. 17e, the ultimate voltage of the predicted waveform was 0.13, while the ultimate voltage of the real waveform was − 0.02. In Fig. 17f, the ultimate voltage of the predicted waveform was 0.28, while the ultimate voltage of the real waveform was − 0.30. This is because soil may contain contaminants or ionic compounds, leading to excessively high electrical conductivity. However, such contamination information cannot be characterized in CI curves, making it challenging for TDR waveforms generated from CI curves to accurately predict real electrical conductivity.

7 Discussions

7.1 Inherent correlations of soil multi-source heterogeneous data

Based on the model testing results in Chapter 6, the inherent correlations of the three types of soil multi-source heterogeneous data are illustrated in Fig. 18. Since the CI can be generated from RGB images, it indicates that cone index information is embedded in soil images. However, RGB images cannot be generated from CI, while binary images, which retain only soil texture after removing color information, can be generated from CI. It suggests that CI contains texture information about the soil but does not include color-related information, such as mineral content.

The permittivity-related information in TDR waveforms can be inferred through RGB images, but electrical conductivity-related information in TDR waveforms cannot be accurately characterized by RGB images. It indicates that soil images and TDR waveforms share overlapping information related to soil water content, but soil images are insufficient to directly represent the presence or concentration of pollutants such as ionic compounds. RGB images cannot be generated from TDR waveforms, but binary images, which retain only texture information, can be generated from TDR waveforms. It implies that TDR waveforms contain texture information about the soil but not color-related information, such as mineral content.

Since CI can be generated from TDR waveforms, it indicates that TDR waveforms contain significant information that of cone index. The permittivity-related information in TDR waveforms can also be converted into CI, while electrical conductivity-related information in TDR

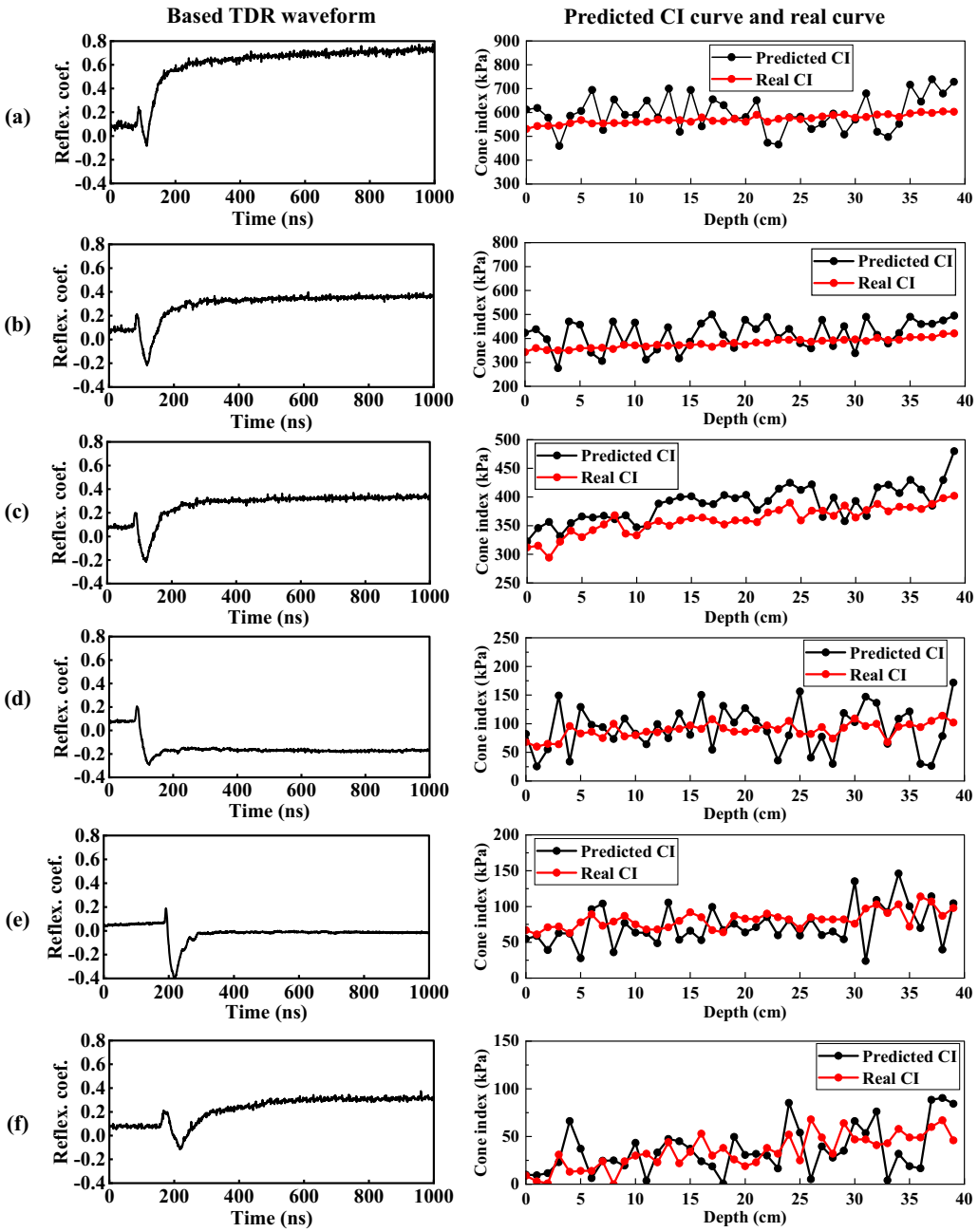


Fig. 16 Generated CI curve based on TDR waveform

waveforms cannot be accurately represented by CI. It indicates that cone index and TDR waveforms share overlapping information regarding soil water content, but the cone index cannot represent the presence or concentration of pollutants such as ionic compounds.

In summary, the inherent correlations of the three multi-source heterogeneous soil indices in different dimensions can be illustrated as shown in Fig. 18.

7.2 Pros and cons of the correlation analysis method

The generative-model-aided correlation analysis method soil properties proposed in this work, aims to address the challenges of dimensional matching and information loss encountered during the correlation analysis of multi-dimensional raw data. This method introduces a novel methodology to analyzing property correlations using neural networks. By applying this method, indices containing extensive raw data with coupling relationship no

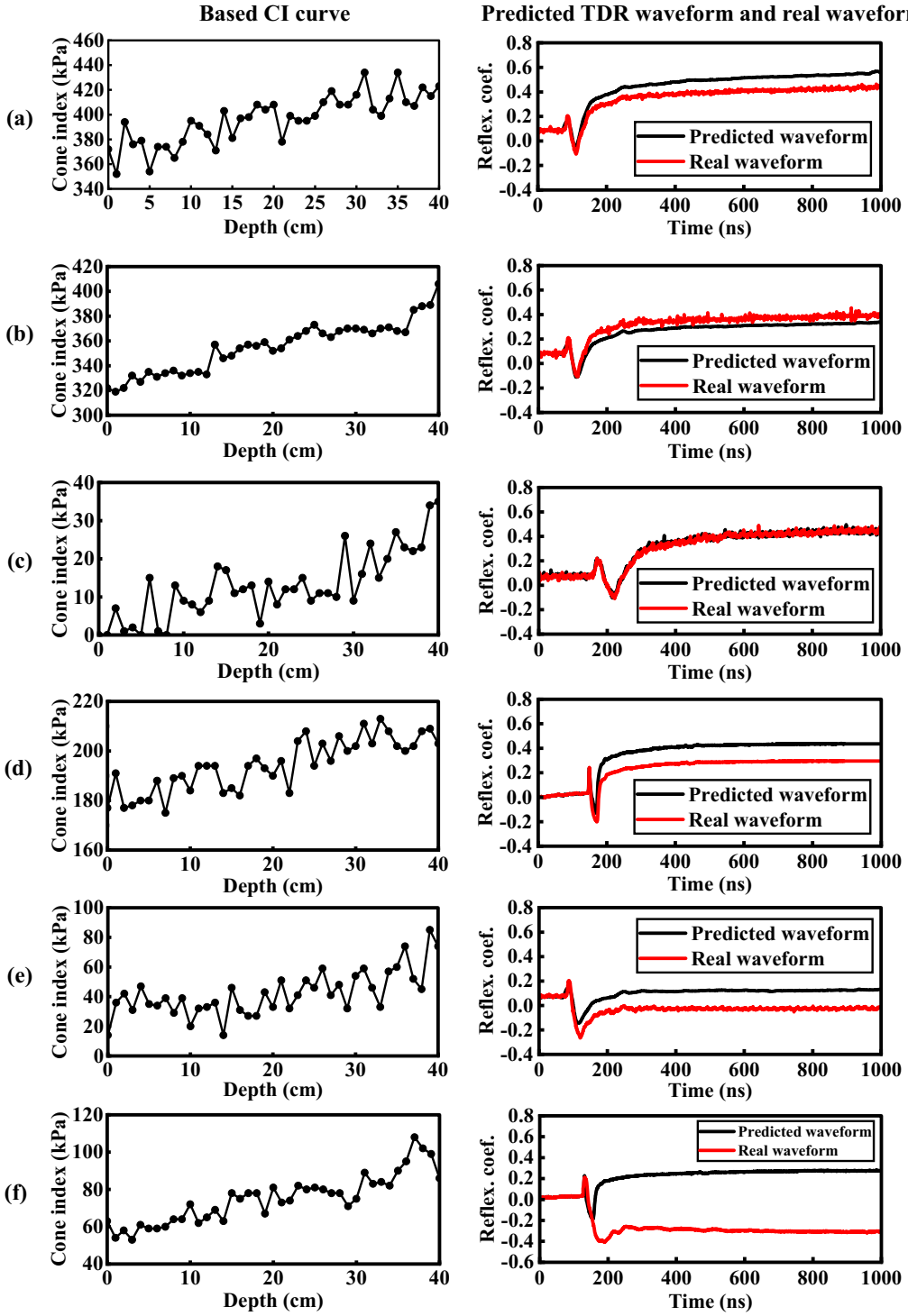


Fig. 17 Generated TDR waveform based on CI curve

longer require feature extraction or information integration for further analysis, thereby avoiding information loss. In addition, leveraging the characteristics of generative models, this method enables the correlation analysis of indices across one-dimensional, two-dimensional, and multi-dimensional spaces without the need for dimensional normalization.

As for applicability, it can be observed that for coarse-grained soils, the generated soil RGB images exhibit higher similarity to real soil RGB images. In contrast, for fine-grained soils, the generated RGB images often display

Indices	Case	Results
Soil image & cone index	img → CI	CI can be generated based on RGB images
	CI → img	RGB images cannot be generated based on CI, but binary images can be generated based on CI.
Soil image & TDR waveform	img → TDR	Permittivity related info in the TDR waveform can be generated based on RGB images, but the EC related info cannot be generated.
	TDR → img	RGB images cannot be generated based on TDR waveforms, but binary images can be generated based on TDR waveforms.
Cone index & TDR waveform	CI → TDR	CI can be generated based on TDR waveforms, but not stable.
	TDR → CI	Permittivity related info in the TDR waveform can be generated based on CI, but the EC related info cannot be generated.

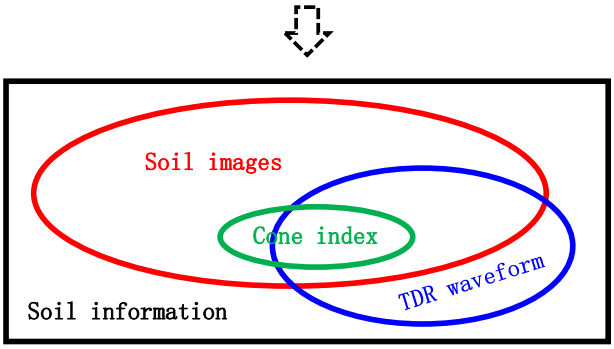


Fig. 18 Inherent correlations of multi-source heterogeneous data

overexposed white spots, leading to a loss of image information and impairing the expressive quality of the generated results. For example, in Fig. 10, the coarse-grained samples (a–d) exhibit a stronger granular texture in the generated soil RGB images, whereas the fine-grained clay samples (g and h) show excessive overexposed highlights in the generated RGB images, compromising the color representation of the soil. Similarly, in Fig. 13, the coarse-grained samples (a–c) also demonstrate more significant granularity in the generated soil RGB images. In contrast, the fine-grained samples (d–h) all exhibit overexposure, with the overexposure becoming more noticeable as the particle size decreases and the water content increases, ultimately leading to color distortion in the generated clay RGB images. Regarding the generated cone index, particle size seems not appear to have a significant impact on the experimental results. For the generated TDR waveforms, fine-grained samples seem more challenging to generate accurately. As shown in Fig. 17, the TDR waveform generation results for samples d–f deviate considerably from the real measurements. However, fine-grained soils, due to their large specific surface area and cation exchange capacity, tend to adsorb more free ions, leading to inaccuracies in the generated TDR waveforms. In summary, the coarser the soil particles, the closer the generated indices are to the real measurements, containing more detailed information and resulting in better model performance.

However, this method currently has several limitations. First, while retaining a large amount of raw data avoids data loss, it introduces a significant amount of noise and irrelevant information without physical significance. This increases the difficulty of correlation analysis and may affect accuracy. Second, when comparing the performance metrics of trained models across different dimensions, the lack of consistent performance metrics makes it challenging to establish a unified evaluation standard. Third, the inherent correlations between different indices are complex and diverse, including containment, intersection, and complementation. At present, this method is more effective at handling containment correlations, but it struggles to quantitatively address intersecting relationships. Further research on this method will focus on improving the evaluation framework and exploring quantitative methods for analyzing complex relationships.

The dataset supporting this study primarily consists of fine-grained soil samples. Therefore, the findings of this research are more applicable to silts and clays. Given the substantial sample size of 23,122, the extensive dataset ensures strong generalizability of the results. However, due to varying natural environmental influences across different regions, soils exhibit distinct mineralogical compositions—such as the black soils in Northeast China, collapsible loess in Northwest China, and red clays in Southwest China. Since the database does not encompass all possible soil types, the research results inevitably have some limitations.

To address above limitations, some measures can be implemented: (1) Advanced Preprocessing Pipelines: implementing wavelet transform-based denoising for TDR waveforms to maintain signal integrity while eliminating high-frequency noise, combined with attention-guided masking algorithms for soil images to automatically suppress non-soil background interference; (2) Physics-Informed Data Augmentation: generating physically realistic synthetic training data through discrete element modeling of cone penetration tests across varied soil compositions and stochastic process simulation of TDR responses governed by Maxwell's equations; (3) Hybrid Model Architecture: replacing purely data-driven DDPM with multimodal gating mechanisms that dynamically adjust sensor contributions, while integrating uncertainty quantification modules to identify and flag predictions compromised by input noise. This integrated approach synergistically enhances data quality, expands training diversity with physical constraints, and improves model robustness through principled architectural innovations.

8 Conclusions

Excavated soil, as a type of sustainable geomaterial characterized by large production, low environment threat, and huge resource potential, requires a comprehensive understanding of engineering, environmental, and resource properties for its utilization. In this work, an excavated soil information collecting system (ESICS) was developed to gather 3243 groups of multi-source heterogeneous data, including soil RGB images, cone index curves, and TDR waveforms. After data augmentation, a big database containing 23,122 sets of data on soil surface morphology, mechanical properties, and electrical properties was established. Then, a generative-model-aided correlation analysis method of soil properties was proposed. The inherent correlations of soil indices across different dimensions in the database were investigated. The key innovations are as follows:

1. **Soil database of surface morphology, mechanical properties, and electrical properties:** A multi-source heterogeneous database containing 23,122 sets of data was created, including soil images (2D data), cone index curves (1D spatial-series data), and TDR waveforms (1D time-series data). This database characterizes the engineering, environmental, and resource properties of excavated soils.
2. **Generative-model-aided correlation analysis method of soil properties:** By using generative deep learning models as research tools, data generation tests (forward and reverse tests) are implemented between pairs of indices with unknown relationships. The performance metrics of the generative models are compared and analyzed to determine the correlations among the unknown indices. This method is suitable for analyzing correlations across different dimensions involving large amounts of raw data, thereby avoiding data loss.
3. **Inherent correlations of soil images, cone index, and TDR waveforms:** Using the database mentioned in (1) as a case study, the cutting-edge generative model—diffusion model—was employed to explore the inherent correlations of the multi-source heterogeneous data of excavated soils. The results revealed that there is overlapping information between soil images and TDR waveforms, and the cone index information is contained within both soil images and TDR waveforms.

Future research directions are proposed to enhance the applicability and robustness of the proposed method. First, the framework's application scope could be expanded to environmental engineering and soil science domains through the integration of soil environmental indicators and advanced characterization data, enabling the development of comprehensive predictive models. Second, the multimodal data system could be strengthened by incorporating 3D microstructural datasets and designing adaptive fusion algorithms for heterogeneous data types. Third, the deep learning architecture could be optimized through domain-specific neural network designs and targeted learning strategies to improve handling of multi-source soil data. Finally, the critical transition from laboratory research to field applications could be facilitated by validating the model under real-world engineering conditions and developing noise-robust adaptive algorithms for field data processing. These proposed extensions would significantly advance both the scientific depth and practical utility of the method while maintaining its rigorous theoretical foundation.

Acknowledgements This work was supported by Basic Science Center Program for Multiphase Evolution in Hypergravity of the National Natural Science Foundation of China (No. 51988101), Natural Science Foundation of Hunan Province—a cooperation with China Construction Fifth Engineering Division Corp., Ltd (2023JJ70027), Academic Star Training Program for Ph.D. Students of Zhejiang University (No. 2022045), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 15220221, 15227923, 15229223), and the Research Centre for Resources Engineering toward Carbon Neutrality (RCRE) of The Hong Kong Polytechnic University (No. 1-BBEM). We gratefully thank the funding of Zhejiang Lvnong Ecological Environment Co., Ltd. and Shenergy Environment Co., Ltd.

Authors' contributions Q-MG, Z-YY, and L-T were involved in conceptualization; Q-MG helped in methodology; Q-M and L-TZ contributed to system design and building; Q-MG, HF, G-QY, and Y-AC were involved in formal analysis and investigation; Q-MG helped in writing—original draft preparation; Z-YY and L-TZ helped in writing—review and editing; Z-YY, L-TZ, and Y-MC contributed to resources.

Funding Open access funding provided by The Hong Kong Polytechnic University.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethics approval and consent to participate This is nonhuman subject research and waived the need for informed consent.

Consent to participate/publish All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References

1. Ahmad Z, Jaffri ZuA, Chen M, Bao S (2024) Understanding GANs: fundamentals, variants, training challenges, applications, and open problems. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-19361-y
2. Alessandrini M, Falaschetti L, Biagetti G, Crippa P, Luzzi S, Turchetti C (2023) A deep learning model for correlation analysis between electroencephalography signal and speech stimuli. Sensors 23:8039
3. Ameratunga J, Sivakugan N, Das BM (2016) Correlations of soil and rock properties in geotechnical engineering. Springer, City
4. Asghari V, Leung YF, Hsu S-C (2020) Deep neural network based framework for complex correlations in engineering metrics. Adv Eng Inform 44:101058. https://doi.org/10.1016/j.aei.2020.101058
5. Ayers P, Bozdech G, Freeman J, Reid A, O'Kins J (2011) Development of a dynamic visco-elastic vehicle-soil interaction model for rut depth, energy and power determinations. In: 17th international conference of the international society for terrain-vehicle systems, City
6. Azqadan E, Jahed H, Arami A (2023) Predictive microstructure image generation using denoising diffusion probabilistic models. Acta Mater 261:119406. https://doi.org/10.1016/j.actamat.2023.119406
7. Bai H, Zhou X, Zhao Y, Zhao Y, Han Q (2023) Soil CT image quality enhancement via an improved super-resolution reconstruction method based on GAN. Comput Electron Agric 213:108177. https://doi.org/10.1016/j.compag.2023.108177
8. Bharambe U, Mahato M, Durbha S, Dhavale C (2024) Exploring opportunities of generative artificial intelligence for sustainable soil analytics in agriculture. In: Sharma C, Shukla AK, Pathak S, Singh VP (eds) Sustainable development and geospatial technology: applications and future directions, vol 2. Springer Nature, Switzerland City, pp 23–43
9. Bobko P (2001) Correlation and regression: Applications for industrial organizational psychology and management. Sage, City
10. Bonett DG (2008) Meta-analytic interval estimation for bivariate correlations. Psychol Methods 13:173
11. Carter M, Bentley SP (1991) Correlations of soil properties. Pentech Press, City
12. Carter M, Bentley S (2016) Soil properties and their correlations, 2nd edn. City
13. Chen PY, Popovich PM (2002) Correlation: parametric and nonparametric measures. Sage, City
14. Chen W, Ding J, Wang T, Connolly DP, Wan X (2023) Soil property recovery from incomplete in-situ geotechnical test data using a hybrid deep generative framework. Eng Geol 326:107332. https://doi.org/10.1016/j.enggeo.2023.107332
15. Ching J, Lin G-H, Phoon K-K, Chen J (2017) Correlations among some parameters of coarse-grained soils—the multivariate probability distribution model. Can Geotech J 54:1203–1220. https://doi.org/10.1139/cgj-2016-0571
16. Cristóbal J, Foster G, Caro D, Yunta F, Manfredi S, Tonini D (2024) Management of excavated soil and dredging spoil waste from construction and demolition within the EU: practices, impacts and perspectives. Sci Total Environ 944:173859. https://doi.org/10.1016/j.scitotenv.2024.173859
17. ElMouchi A, Siddiqua S, Wijewickreme D, Polinder H (2021) A review to develop new correlations for geotechnical properties of organic soils. Geotech Geol Eng 39:3315–3336. https://doi.org/10.1007/s10706-021-01723-0
18. European Commission (2024) Excavated soil generation, treatment and reuse in the EU—Final report for Task 1.1 of the support study for implementing the EU Soil Strategy for 2030 (09.0201/2022/877182/SER/D.1). Publications Office of the European Union, City
19. Fan G, Zhang N, Lv S, Cabrera MB, Yuan J, Fan X, Liu H (2022) Correlation analysis of chemical components and rheological properties of asphalt after aging and rejuvenation. J Mater Civ Eng 34:04022303. https://doi.org/10.1061/(ASCE)MT.1943-5533.0004467
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63:139–144. https://doi.org/10.48550/arXiv.1406.2661
21. Guo Q-M, Zhan L-T, Yin Z-Y, Feng H, Yang G-Q, Chen Y-M (2024) Multi-modal fusion deep learning model for excavated soil heterogeneous data with efficient classification. Comput Geotech 175:106697. https://doi.org/10.1016/j.compgeo.2024.106697
22. Guo Q, Zhan L, Shen Y, Wu L, Chen Y (2022) Classification and quantification of excavated soil and construction sludge: a case study in Wenzhou, China. Front Struct Civ Eng 16:202–213. https://doi.org/10.1007/s11709-021-0795-8
23. Han D, Xiong W, Chen Y, Xu J (2023) Mechanical properties of excavated soil waste-based cementitious products at normal condition or after water soaked: a literature review of experimental results. Case Stud Construct Mater 18:e02111. https://doi.org/10.1016/j.cscm.2023.e02111
24. He S-f, Shen L-m, Xie H-x (2021) Hyperspectral estimation model of soil organic matter content using generative adversarial networks. Spectrosc Spectr Anal 41:1905–1911. https://doi.org/10.3964/j.issn.1000-0593(2021)06-1905-07

25. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Advances in neural information processing systems, vol 33, pp 6840–6851. https://doi.org/10.48550/arXiv.2006.11239
26. Jayalekshmi S, Elamathi V (2020) A review on correlations for consolidation characteristics of various soils. IOP Conf Ser Mater Sci Eng 1006:012007. https://doi.org/10.1088/1757-899X/1006/1/012007
27. Karabulut EÖ, Koyuncu M (2007) Neural network-based correlations for the thermal conductivity of propane. Fluid Phase Equilib 257:6–17. https://doi.org/10.1016/j.fluid.2007.04.024
28. Karyati K, Ipor IB, Jusoh I, Wasli ME (2018) Correlation between soil physicochemical properties and vegetation parameters in secondary tropical forest in Sabal, Sarawak, Malaysia. IOP Conf Ser Earth Environ Sci 144:012060. https://doi.org/10.1088/1755-1315/144/1/012060
29. Kashyap VS, Sancheti G, Yadav JS, Agrawal U (2023) Smart sustainable concrete: enhancing the strength and durability with nano silica. Smart Construct Sustain Cities 1:20. https://doi.org/10.1007/s44268-023-00023-1
30. Kingma DP (2013) Auto-encoding variational bayes. https://doi.org/10.48550/arXiv.1312.6114
31. Kumar V (2024) Estimation of spatial variability of soil profile using hydraulic operated real time dual sensor based integrated system. J Inst Eng India Ser A 105:431–445. https://doi.org/10.1007/s40030-024-00802-8
32. Liu S, Zou H, Cai G, Bheemasetti TV, Puppala AJ, Lin J (2016) Multivariate correlation among resilient modulus and cone penetration test parameters of cohesive subgrade soils. Eng Geol 209:128–142. https://doi.org/10.1016/j.enggeo.2016.05.018
33. Lo Man K, Loh Daniel RD, Chian Siau C, Ku T (2023) Probabilistic prediction of consolidation settlement and pore water pressure using variational autoencoder neural network. J Geotech Geoenviron Eng 149:04022119. https://doi.org/10.1061/JGGEFK.GTENG-10555
34. Lu X, Yang Q, Wang H, Zhu Y (2023) A global meta-analysis of the correlation between soil physicochemical properties and lead bioaccessibility. J Hazard Mater 453:131440. https://doi.org/10.1016/j.jhazmat.2023.131440
35. Luleci F, Catbas FN (2023) A brief introductory review to deep generative models for civil structural health monitoring. AI Civ Eng 2:9. https://doi.org/10.1007/s43503-023-00017-z
36. Meimaroglou N, Mouzakis C (2024) The role of intrinsic soil properties in the compressive strength and volume change behavior of unstabilized earth mortars. Mater Struct 57:50. https://doi.org/10.1617/s11527-024-02314-0
37. Meng C, Yang W, Wang D, Hao Z, Li M (2023) Shadow removal method of soil surface image based on GAN used for estimation of farmland soil moisture content. Meas Sci Technol 34:085114. https://doi.org/10.1088/1361-6501/acd133
38. Michon L, Hanquet B, Diawara B, Martin D, Planche J-P (1997) Asphalt study by neuronal networks. correlation between chemical and rheological properties. Energy Fuels 11:1188–1193. https://doi.org/10.1021/ef9700386
39. Olivares BO, Lobo D, Carlos Rey J, Vega A, Rueda MA (2023) Relationships between the visual evaluation of soil structure (VESS) and soil properties in agriculture: a meta-analysis. Sci Agropecuaria 14:67–78. https://doi.org/10.17268/sci.agropecu.2023.007
40. Pan S, Abouei E, Wynne J, Chang CW, Wang T, Qiu RL, Li Y, Peng J, Roper J, Patel P (2024) Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. Med Phys 51:2538–2548. https://doi.org/10.48550/arXiv.2305.19467
41. Paul SC, Basit MA, Hasan NMS, Islam MS (2024) Sustainable cement mortar production using rice husk and eggshell powder: a study of strength, electrical resistivity, and microstructure. Smart Construct Sustain Cities 2:13. https://doi.org/10.1007/s44268-024-00037-3
42. Peng J, Qiu RL, Wynne JF, Chang CW, Pan S, Wang T, Roper J, Liu T, Patel PR, Yu DS (2024) CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model. Med Phys 51:1847–1859. https://doi.org/10.1002/mp.16704
43. Roy T, Das P, Jagirdar R, Shhabat M, Abdullah MS, Kashem A, Rahman R (2025) Prediction of mechanical properties of eco-friendly concrete using machine learning algorithms and partial dependence plot analysis. Smart Construct Sustain Cities 3:2. https://doi.org/10.1007/s44268-025-00048-8
44. Sainju UM, Liptzin D (2022) Relating soil chemical properties to other soil properties and dryland crop production. Front Environ Sci. https://doi.org/10.3389/fenvs.2022.1005114
45. Sainju UM, Liptzin D, Jabro JD (2022) Relating soil physical properties to other soil properties and crop yields. Sci Rep 12:22025. https://doi.org/10.1038/s41598-022-26619-8
46. Samuel-Rosa A, Dalmolin RSD, Miguel P, Zalamena J, Dick DP (2013) The effect of intrinsic soil properties on soil quality assessments. Rev Bras Ciênc Solo. https://doi.org/10.1590/S0100-06832013000500013
47. Shoemaker Travis A, Beaino C, Centella RDM, Zhao W, Tanissa C, Lawrence J, Hashash Youssef MA (2023) Generative AI: the new geotechnical assistant? J Geotech Geoenviron Eng 149:02823004. https://doi.org/10.1061/JGGEFK.GTENG-11859
48. Strechan AA, Kabo GJ, Paulechka YU (2006) The correlations of the enthalpy of vaporization and the surface tension of molecular liquids. Fluid Phase Equilib 250:125–130. https://doi.org/10.1016/j.fluid.2006.10.007
49. Teixeira F, Basch G (2019) Correlation between VSA soil indicators and measured soil properties. In: Performance of promising land management practices to populate recommendations of SQAPP iSQAPER project deliverable, vol 6, pp 45
50. Tsimpouris E, Tsakiridis NL, Theocharis JB (2021) Using autoencoders to compress soil VNIR–SWIR spectra for more robust prediction of soil properties. Geoderma 393:114967. https://doi.org/10.1016/j.geoderma.2021.114967
51. Varsha Pandey PG, Singh AP (2019) Correlation between physical, chemical and biological properties of soil under different land use systems. Int J Chem Stud 7:469–471
52. Verbrugge JC, Schroeder C (2018) Geotechnical correlations for soils and rocks. Wiley, City
53. Wang C, Wei M, Zhang J, Li X, Liu J (2024) Reconstructing in-situ soil monitoring data using variational autoencoder-based generative compressed sensing. https://ssrn.com/abstract=4924885
54. Yan T, Shen S-L, Zhou A (2025) Data augmentation-assisted muck image recognition during shield tunnelling. Undergr Space 21:370–383. https://doi.org/10.1016/j.undsp.2024.10.001
55. Zhan L, Guo Q, Mu Q, Chen Y (2021) Detection of ionic contaminants in unsaturated soils using time domain reflectometry penetrometer. Environ Earth Sci 80:330. https://doi.org/10.1007/s12665-021-09618-2
56. Zhan L-t, Guo Q-m, Chen Y-m, Wang S-y, Feng T, Bian Y, Wu J-j, Yin Z-y (2023) An efficient classification system for excavated soils using soil image deep learning and TDR cone penetration test. Comput Geotech 155:105207. https://doi.org/10.1016/j.compgeo.2022.105207
57. Zhan L, Wang Z, Deng Y, Zeng Q, Chen P, Chen Y (2024) Hydrothermal solidification of siliceous and calcareous wastes into building materials: a generic mix design framework. Dev Built Environ. https://doi.org/10.1016/j.dibe.2024.100534
58. Zhang Y, Evans JRG, Yang S (2020) Exploring correlations between properties using artificial neural networks. Metall Mater Trans A 51:58–75. https://doi.org/10.1007/s11661-019-05502-8

59. Zhang L, Qiu Y, Wu T, Zhang W (2023) Correlation analysis of physical and mechanical parameters of inland fluvial-lacustrine soft soil based on different survey techniques. Appl Rheol. https://doi.org/10.1515/arh-2022-0145
60. Zhao W-W, Shen S-L, Yan T, Zhou A (2024) Intelligent approach for mucky soil identification during shield tunnelling by enhanced YOLO model. J Rock Mech Geotech Eng. https://doi.org/10.1016/j.jrmge.2024.09.025