Computer-Aided Civil and Infrastructure Engineering 51 (2026) 100170

Contents lists available at ScienceDirect

Computer-Aided Civil and Infrastructure Engineering

journal homepage: www.sciencedirect.com/journal/computer-aided-civil-and-infrastructure-engineering

Enhancement of intelligent GPR detection based on category-aware generative framework of B-scan images

Zuhao Wang [a], Guangyao Li [a,b], Fengzijun Pan [a,c], Zepeng Fan [a,c,*], Xiaopei Cai [d], Dawei Wang [a,c]

[a] School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, China
[b] Faculty of Civil Engineering, RWTH Aachen University, Aachen, Germany
[c] Chongqing Research Institute, Harbin Institute of Technology, Chongqing, China
[d] School of Civil Engineering, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Keywords:
Ground penetrating radar
Intelligent detection
Data augmentation
Class imbalance
Rectified flow model

ABSTRACT

Deep learning has advanced intelligent detection of ground penetrating radar (GPR) applied in transportation infrastructure inspection, yet its performance remains limited by the scarcity and class imbalance of labeled GPR B-scan images. To address this challenge, this study proposes GPR-ImaGen, a category-aware generative data augmentation framework for GPR B-scan images. Built on the pretrained rectified flow model Flux.1-dev, the framework generates defect category-consistent B-scan images under semantic guidance from engineering annotations, alleviating the scarcity and class imbalance of the training data for intelligent detection. To better adapt the model to the characteristics of GPR B-scan images and limited-data conditions, parameter-efficient fine-tuning, a robust loss function and a boundary-focused timestep sampling strategy are incorporated into the training process. The proposed framework achieves an FID of 10.496 and a Bscan-FID of 10.828. Extensive task-oriented experiments across multiple intelligent detection models and diverse limited-data settings show that the generated data consistently improve downstream detection performance. Among them, YOLOv11 achieves the highest detection performance (mAP@50 of 91.6%) at a 100% augmentation ratio, while mAP@50 increases from 73.6% to 81.9% under a data-scarce setting with only 25% of the real training data retained. These results indicate the practical potential of a category-aware generative framework for improving the robustness and reliability of intelligent GPR detection for transportation infrastructure inspection.

1. Introduction

Ground penetrating radar (GPR) has become a widely used non-destructive technique for assessing subsurface conditions in transportation infrastructure because of its high efficiency, continuous data acquisition capability, and sensitivity to internal structural anomalies (Cui et al., 2025; Wang et al., 2025; Wang et al., 2023). However, GPR measurements are typically represented as B-scan images that contain complex electromagnetic reflection patterns rather than direct structural images, making reliable interpretation difficult in practical inspection scenarios. In recent years, deep learning has significantly advanced the intelligent interpretation of GPR B-scan images, with the use of both two-stage detectors such as Faster R-CNN (Pham and Lefevre, 2018) and Mask R-CNN (Qin et al., 2021), and one-stage detectors such as the YOLO series (Li et al., 2023a). These methods have substantially improved detection efficiency and accuracy, making large-scale intelligent GPR inspection more feasible (Lv et al., 2025). Nevertheless, the performance of such models still depends strongly on the availability and quality of labeled training data, and data scarcity remains a key factor limiting the practical application of deep learning-based intelligent GPR detection (Xiong et al., 2024).

In practical GPR applications, acquiring high-quality annotated dataset remains costly and labor-intensive because subsurface defects are not directly visible and usually require destructive validation or additional sensing to establish ground truth (Yue et al., 2022; Li et al., 2024). As a result, available GPR B-scan datasets for intelligent defect detection are typically limited to hundreds or a few thousands annotated images, which is substantially smaller than the scale of general computer vision benchmarks used for training modern defect models. Recent pavement crack detection studies, for instance, have used datasets containing thousands of images, including a double-labeled pavement crack dataset with 20,000 images (Li et al., 2023b). In contrast, the

* Corresponding author at: School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, China.
E-mail address: zepeng.fan@hit.edu.cn (Z. Fan).

https://doi.org/10.1016/j.cacaie.2026.100170
Received 19 April 2026; Received in revised form 5 July 2026; Accepted 3 August 2026
Available online 4 August 2026
1093-9687/© 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

recent YOLO-based GPR detection study collected only 812 original real GPR images from six roads (Li et al., 2023a). This contrast indicates that intelligent GPR detection is often developed under more pronounced real-data scarcity. Moreover, real-world GPR dataset often shows severe class imbalance and limited coverage of rare but critical defect categories. To address this problem, data augmentation has been widely adopted. Conventional augmentation methods, such as geometric transformations and intensity perturbations, can increase the number of training samples. But these methods mainly introduce local variations of existing images and provide limited new structural information. To further enrich the data distribution, domain knowledge-based methods have been used to simulate variations in acquisition parameters (Yue et al., 2022). In addition, physics-based forward simulations, such as finite-difference time-domain (FDTD) modeling (Feng et al., 2018), can generate synthetic GPR B-scan images in a controllable manner. However, these simulated data often fail to capture the complex noise characteristics and environmental variability of real-world measurements, resulting in a noticeable domain gap between simulated and real GPR data and limiting the direct transferability of deep learning-based models to practical applications (Zhang et al., 2025).

More recently, deep learning-based generative models have emerged as a promising approach for distribution-level data augmentation in GPR B-scan image analysis (Xiong et al., 2024). Unlike conventional augmentation methods that operate through deterministic transformations, these models aim to learn the underlying data distribution and generate new samples that preserve both the structural patterns and statistical characteristics of real B-scan images. Among them, generative adversarial networks (GANs) have been widely explored for GPR data generation. Qin et al. combined FDTD simulations with DCGAN-based image generation for data augmentation in tunnel lining inspection, showing that synthetic GPR images can improve automatic recognition performance when measured data are limited (Qin et al., 2021). Yao et al. combined gprMax-based numerical simulation with GAN-generated GPR images to augment datasets for subsurface road target detection, and demonstrated that the augmented data improved detection accuracy (Yao, 2025). Zhao et al. proposed WAEGAN, which introduced a Wasserstein loss to improve training stability and a pretrained classifier to impose category constraints, thereby enabling multi-category B-scan data generation (Gulrajani, 2017; Zhao et al., 2023). Despite these advances, the adversarial training process remains sensitive to optimization conditions, and category control often relies on external supervision rather than intrinsic semantic conditioning. This restricts their ability to capture complex radar signal structures and support robust category-aware GPR B-scan image generation under class-imbalanced conditions (Zhao et al., 2023).

Along with GAN-based studies, diffusion-based generative models have also been introduced for GPR B-scan image generation. As a representative class of generative models, denoising diffusion probabilistic models (DDPMs) progressively transform random noise into structured images through iterative denoising, providing a stable training framework and strong capacity for modeling complex data distributions (Ho et al., 2020). These properties make diffusion models particularly suitable for GPR B-scan images, where radar signals exhibit complex structural patterns and high variability. Chen et al. have shown that diffusion models can generate GPR images with higher fidelity, richer structural details, and improved diversity, which offers a more reliable approach for synthetic dataset construction (Chen et al., 2025a). Furthermore, Li et al. proposed GPR-VMD-Diffusion, which integrates variational mode decomposition with diffusion-based reconstruction to combine simulated structural patterns with real signal features (Li et al., 2025). Their results suggest that diffusion models can improve not only the visual realism of generated radargrams, but also their consistency with real GPR signals in terms of spectral and structural characteristics.

Despite recent progress, existing generative methods for GPR B-scan images still face several critical challenges. First, many current generative frameworks remain dependent on relatively large training dataset and long optimization or sampling processes. Although diffusion models provide improved stability and image fidelity, they still require substantial data support and long iterative procedures to achieve reliable performance. This limits their practicality in the small-sample conditions of GPR applications. Second, controllable generation across defect categories remains limited. Existing methods often rely on external category constraints or primarily focus on signal-level consistency, but provide little intrinsic semantic guidance for targeted generation of category-specific samples. This restricts generative models' capability to alleviate class imbalance issues in engineering applications. Third, the impact of generated data on downstream detection tasks has not been systematically investigated (Chen et al., 2025b). Most previous studies focus on image realism, signal similarity, or limited task-level improvements, while the effectiveness of generative augmentation under different real-data availability, class imbalance, and cross-domain conditions remains insufficiently understood (Tang et al., 2026). These challenges indicate the need for a generative framework that is more data-efficient and semantically controllable, along with a comprehensive evaluation strategy that explicitly examines its practical utility in GPR defect detection. More broadly, recent studies in specialized imaging domains have shown that pretrained vision-language generative foundation models can be adapted using domain-specific image-text pairs to generate semantically consistent images by expert-defined textual descriptions (Bluethgen et al., 2024).

Recent advances in generative modeling have opened a new direction for GPR B-scan generation through rectified flow-based learning. Liu et al. proposed rectified flow as a generative framework based on ordinary differential equation (ODE) that learns distribution transport by encouraging the trajectories to follow straight paths as much as possible (Liu et al., 2023). Trained with a simple least-squares objective, this formulation supports efficient few-step sampling and provides a more direct and computationally attractive alternative to conventional diffusion models based on stochastic differential equations (SDE). More recent studies have further shown that the training dynamics of rectified flow-based models can be strengthened through refining the timestep sampling strategy and loss function (Lee et al., 2024). These characteristics make rectified flow-based models a promising foundation for generation of GPR B-scan images, where efficient sampling and preservation of global reflection geometry are both important.

Motivated by these advantages, this study proposes GPR-ImaGen, a category-aware GPR B-scan image generation framework built on the pretrained rectified flow text-to-image generative model. The pretrained model is adapted to the GPR domain through DoRA (Weight-Decomposed Low-Rank Adaptation)-based parameter-efficient fine-tuning (Liu et al., 2024), while a Pseudo-Huber loss and a U-shaped timestep sampling strategy are introduced to improve training robustness (Lee et al., 2024). In addition, structured text prompts are paired with B-scan images during training to provide semantic guidance for category-aware B-scan image generation. The proposed framework is evaluated through image-quality assessment and downstream detection-oriented data augmentation. Overall, this study aims to examine the capability of rectified flow-based generation to produce realistic GPR B-scan images and its practical utility for intelligent GPR detection.

2. Framework and validation of GPR-ImaGen

This section presents the GPR-ImaGen framework for text-conditioned category-aware generation of GPR B-scan image. As illustrated in Fig. 1, the framework comprises image-text pair construction, a pretrained latent rectified flow backbone, DoRA-based parameter-efficient adaptation, robust flow-matching optimization, and conditional ODE-based generation. The section first introduces the dataset and problem formulation, and then details the main components of the framework, the corresponding generation process, and associated training configuration and computational efficiency assessment. Finally, the methods for validating the superiority of the proposed generative

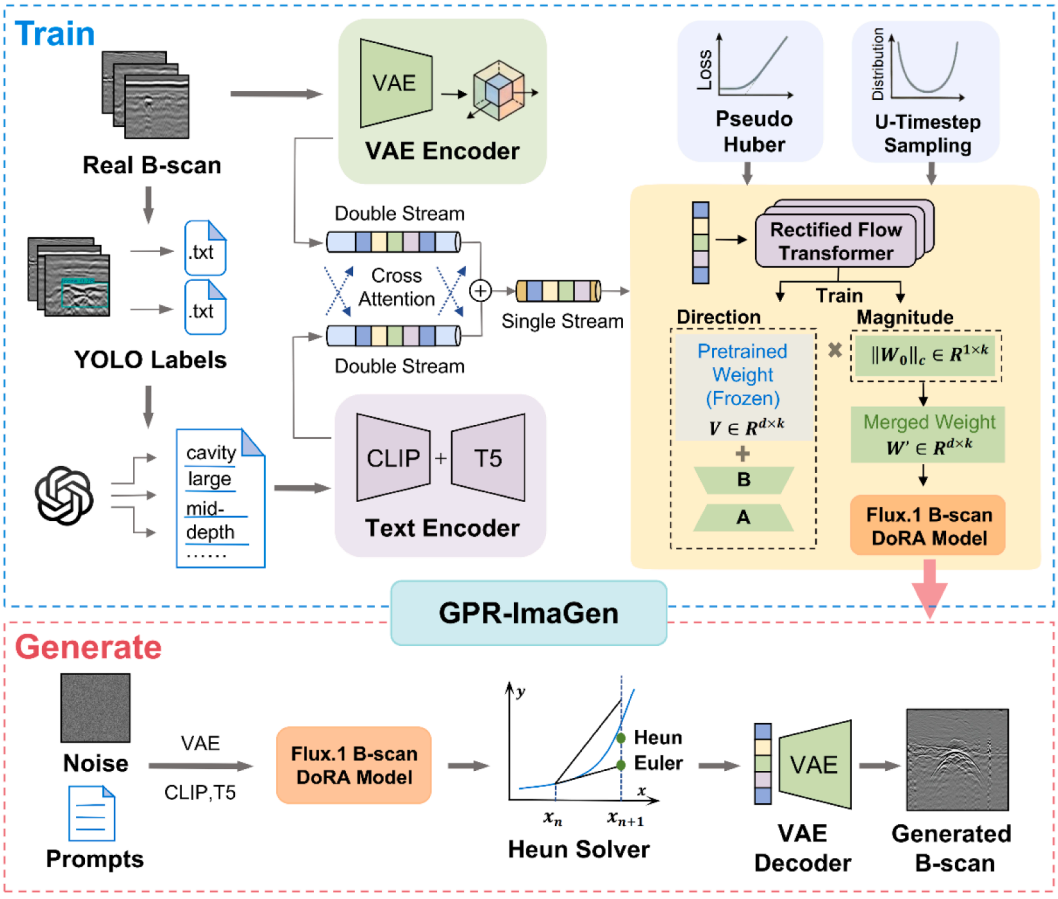


Fig. 1. Framework of GPR-ImaGen. The upper part illustrates the image-text pair construction, VAE encoding, dual-text-encoder conditioning, and DoRA-based rectified flow adaptation described in Sections 2.1–2.2.4. The lower part illustrates the prompt-conditioned ODE-based generation process described in Section 2.2.5.

framework and the effectiveness of the generated data for practical data augmentation in defect detection models are introduced.

2.1. Dataset and problem formulation

This study integrates two publicly available GPR inspection datasets as the source data for category-aware B-scan generation. The first dataset, reported by Mojahid et al. (Mojahid et al., 2025), consists of high-resolution GPR A-scan and B-scan profiles for the intelligent recognition of subsurface cavities, utilities, and intact areas. These data were acquired using 200 MHz and 400 MHz ground-coupled GPR antennas from infrastructure projects conducted in different cities and towns in Morocco between 2019 and 2024. The second dataset is TIGPR, reported by Liu et al. (Liu et al., 2025), which provides multi-view GPR images for transportation infrastructure damage assessment. It covers multiple infrastructure scenarios such as roads, bridges, tunnels, and airports. The key radar acquisition information available for the adopted datasets is summarized in Table 1.

After data consolidation, manual categorization, filtering, and annotation, a total of 873 GPR B-scan images were retained for this study. These images cover six common subsurface defect categories, including cavity, void, interlayer debonding, loose, utility and crack. The retained B-scan images exhibit characteristic GPR reflection patterns, such as hyperbolic signatures generated by subsurface objects, heterogeneous clutter backgrounds, and pronounced speckle noise. While these properties significantly increase the difficulty of intelligent interpretation, they also provide essential structural cues for defect detection. Representative examples of GPR B-scan images from different defect categories are shown in Fig. 2.

For each B-scan image x, the corresponding engineering annotation provides the defect category and approximate spatial location. Based on this metadata, a structured textual condition c is constructed and paired with the image for conditional training (Fan et al., 2025). As shown in Fig. 3, the resulting prompt describes the defect type, approximate location, and scale in a standardized form, thereby establishing an image-text pair for each real GPR B-scan image. In this study, c is automatically generated in batch from engineering annotations by calling the API of large language model (LLM) (GPT-5.2), which converts label information into consistent text descriptions containing defect type, approximate location, and scale. To reduce potential

Table 1
Key information of the adopted open-source GPR datasets.

Dataset	Scenario and configuration	Radar system	Preprocessing	Main categories
Mojahid et al. (Mojahid et al., 2025)	Urban and industrial inspections; ground-coupled survey-wheel acquisition	GSSI ground-coupled antenna with SIR 3000/SIR 4000 control units; 200 and 400 MHz; 512 samples per scan	Time-zero correction, background removal, filtering, and gain processing	Cavity, utility and intact
TIGPR (Liu et al., 2025)	Transportation infrastructure inspection; handheld/vehicle-mounted 2D/3D GPR acquisition; ground-coupled 3D field survey	IDS-FastWave; 2.3 GHz; MALA GX750: 0.75 GHz; GeoScope 3D-Radar: 0.2–3 GHz	Time-zero correction, denoising/filtering, gain adjustment, background removal, and image extraction	Crack, void, loose, and interlayer debonding

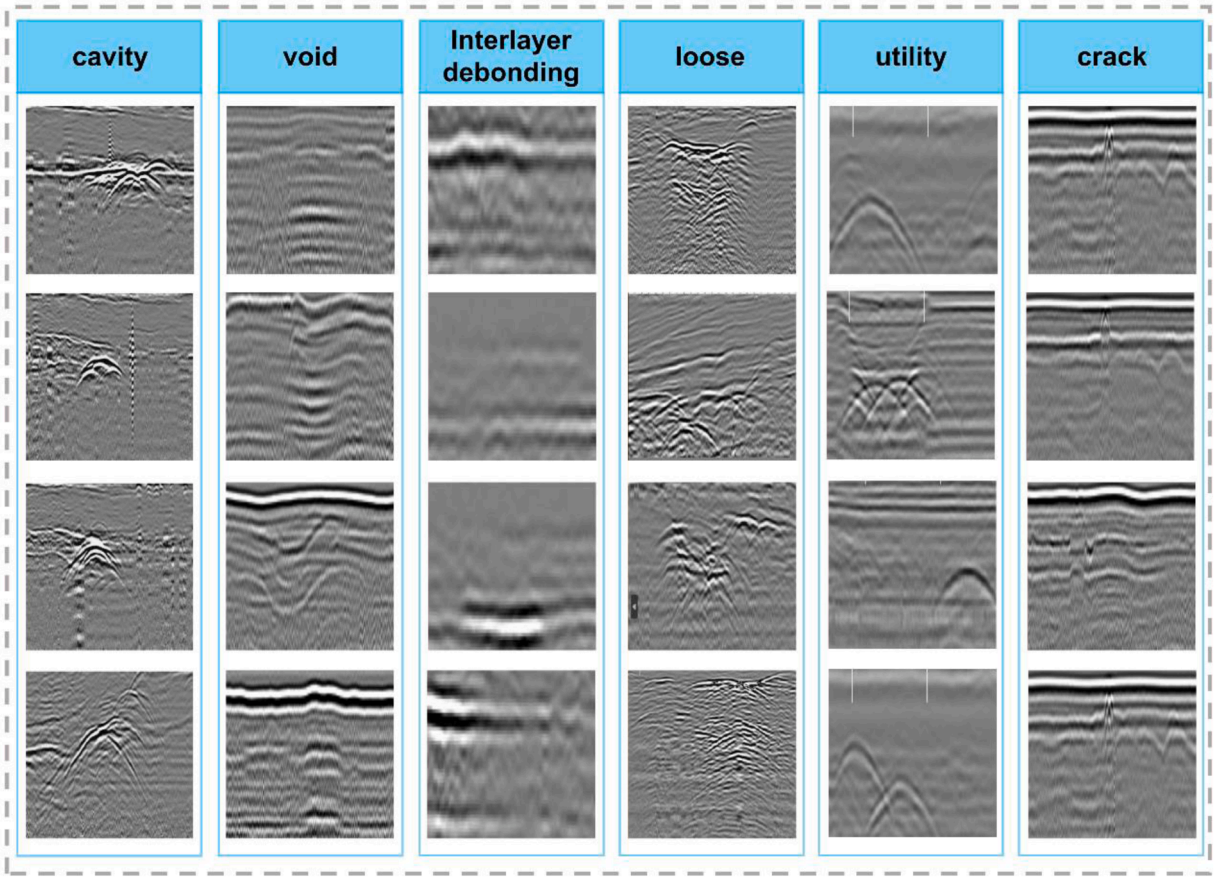


Fig. 2. Representative GPR B-scan images of different defect categories. The examples of cavity and utility are selected from the dataset reported by Mojahid et al. (Mojahid et al., 2025), whereas the examples of void, interlayer debonding, and crack are selected from the TIGPR dataset reported by Liu et al. (Liu et al., 2025). The loose samples were further separated from the cavity samples in the first dataset through manual screening and were used as an independent category in this study.

semantic drift during the language conversion, a representative random subset of the generated prompts is manually audited against the original defect annotations and corresponding B-scan images. The audit confirms that the generated prompts are consistent with the original annotations in terms of defect category, approximate spatial location, and scale, and that no natural-image objects or style-related semantics are introduced. This indicates that the generated prompts preserve the intended engineering semantics of the original annotations and can serve as annotation-consistent textual conditions for model training.

With these quality-controlled image-text pairs, the conditional generation problem is formulated as follows. Formally, let x denote a real GPR B-scan image sampled from the target domain distribution $p_{data}(x)$, and let c denote its paired structured textual condition. Based on these image-text pairs, the objective is to learn a conditional generative model $\widehat{x} \sim p_\theta(x|c)$ that synthesizes B-scan images consistent with the structural characteristics of real B-scan data while enabling category-aware image generation.

2.2. Components of generative framework

2.2.1. Rectified flow backbone

The generative backbone is built on Flux.1-dev, a pretrained text-to-image generative model that formulates image generation as a rectified flow process in the latent space of a variational autoencoder (VAE) (Labs et al., 2025). Instead of directly modeling the generation process in the high-dimensional pixel space, an input image x is first encoded into a compact latent representation z and later reconstructed by the decoder:

$$z = E(x) \tag{1}$$

$$x = D(z) \tag{2}$$

Operating in the latent space significantly reduces computational complexity while preserving high-level structural information, required to represent radar reflections. In the Flux.1-dev architecture, latent image tokens and encoded textual prompts are processed by a transformer backbone that combines double-stream and single-stream blocks with rotary positional embeddings. This design allows textual condition to modulate the latent dynamics while maintaining high computational efficiency.

Following the rectified flow formulation (Liu et al., 2023), the generative process is modeled as a continuous latent trajectory connecting the clean latent representation z and Gaussian noise $\varepsilon \sim N(0, I)$. Specifically, the interpolated latent state at time $t \in [0, 1]$ is defined as:

$$z_t = (1-t)z + t\varepsilon, \varepsilon \sim N(0, I), t \in [0, 1] \tag{3}$$

where t is a continuous time variable controlling the transition from data to noise. When $t = 0$, z_t corresponds to the clean latent representation, while $t = 1$ corresponds to noisy latent state.

Instead of explicitly modeling a discrete diffusion process, the model learns a velocity field $v_\theta(z_t, t, c)$, which transports samples along the latent trajectory under textual conditioning. Accordingly, the training objective is formulated as a flow-matching regression loss, as shown in Eq. (4):

$$L_{flow} = E_{t,z,\varepsilon}\left[\| v_\theta(z_t, t, c) - (\varepsilon - z) \|_2^2\right] \tag{4}$$

This rectified flow formulation enables ODE-based generation and enables efficient few-step numerical integration during inference (Liu et al., 2023). Compared with conventional diffusion models that

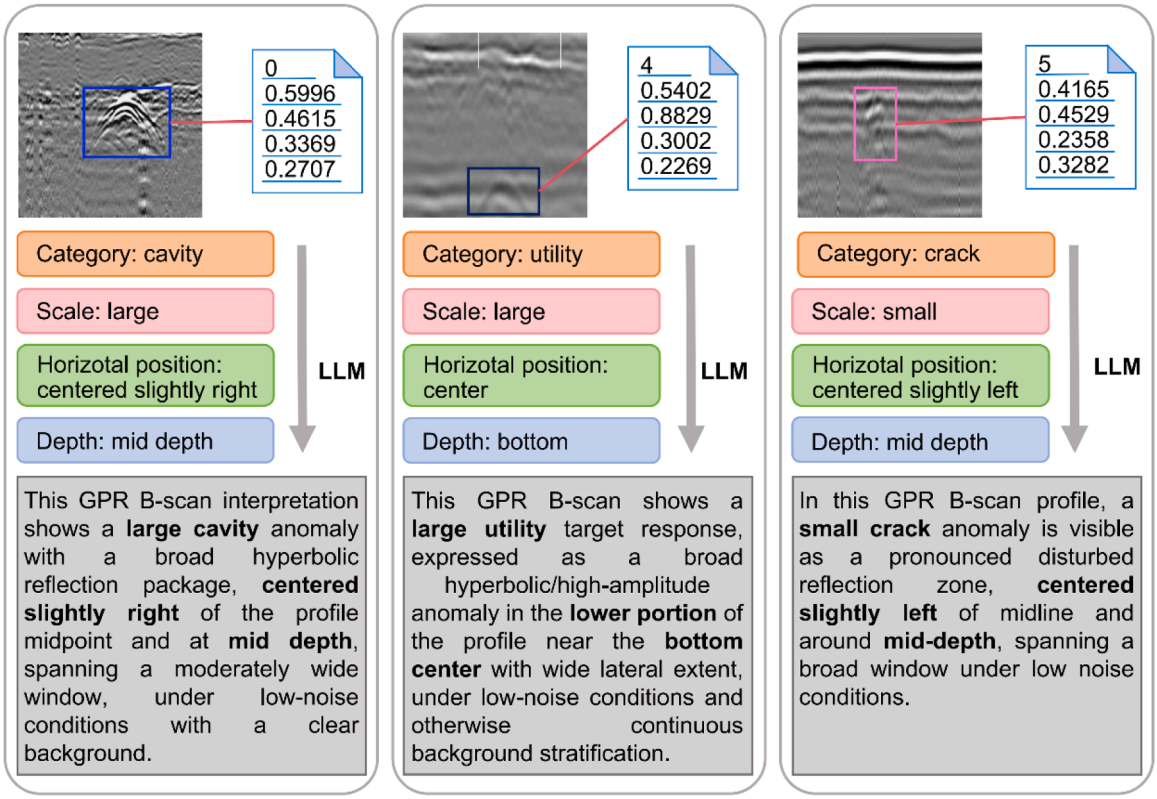


Fig. 3. Image-text pairs for conditional GPR B-scan generation. The five label values beside each B-scan image are YOLO-format annotations, representing the defect class index, normalized center coordinates, and normalized bounding-box width and height.

typically require hundreds of denoising steps, it significantly reduces the number of inference steps while maintaining high generation fidelity. Such efficiency is particularly advantageous for engineering applications that require rapid and controllable GPR B-scan synthesis. Moreover, this backbone provides a general visual prior, which we subsequently adapt to the GPR domain through parameter-efficient fine-tuning.

2.2.2. DoRA-based parameter-efficient fine-tuning

Direct fine-tuning all parameters of a large generative transformer such as Flux.1-dev is computationally expensive and may increase the risk of overfitting when only limited GPR B-scan images are available (Hu et al., 2022). To enable efficient yet expressive domain adaptation, this study adopts DoRA instead of conventional LoRA (Low-Rank Adaptation). Different from standard LoRA, which represents adaptation as a low-rank residual added to the pretrained weight, DoRA reparameterizes each weight matrix into magnitude and direction components and performs adaptation in this decomposed space (Liu et al., 2024). This design is motivated by the observation that the performance gap between LoRA and full fine-tuning is not only related to the number of trainable parameters, but also to the way in which weight updates are represented and optimized. By decoupling magnitude and direction, DoRA yields a learning behavior that is empirically and theoretically closer to full fine-tuning while still preserving the efficiency of parameter-efficient fine-tuning.

Formally, for a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, DoRA expresses it as a column-wise decomposition of magnitude and direction:

$$W_0 = m\frac{V}{\| V \|_c} \tag{5}$$

where $m \in \mathbb{R}^{1\times k}$ is the magnitude vector, $V \in \mathbb{R}^{d\times k}$ is the directional matrix, and $\| \cdot \|$ denotes the vector norm computed for each column. At initialization, m and V are set so that the decomposed representation exactly recovers the pretrained weight. During fine-tuning, the base directional matrix V is kept frozen, whereas the magnitude vector m remains trainable. The directional adaptation is then parameterized by a low-rank update $\Delta V = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and $r \ll \min(d,k)$. Based on the decomposition in Eq. (6), the adapted weight can be written as:

$$W' = m\frac{W_0 + BA}{\| W_0 + BA \|_c} \tag{6}$$

Eq. (6) is more precise than directly adding a low-rank matrix to the full weight, because the update is explicitly constrained to modify the directional component, while the overall scale is controlled separately through m. In addition, A and B follow the standard LoRA-style initialization so that the adapted weight equals the pretrained weight at the start of training. Such decoupling improves optimization flexibility and better preserves the pretrained model prior under limited-data conditions. For Flux.1-dev, DoRA provides a controlled adaptation mechanism that efficiently transfers pretrained visual knowledge to the GPR domain while remaining sensitive to radar-specific structures such as clutter, hyperbolic reflections, and defect-related local responses. In addition, the learned magnitude and directional updates can be merged back into the pretrained weights after training, without introducing extra inference cost.

2.2.3. Robust flow matching with Pseudo-Huber loss

GPR B-scan images often contain strong clutter reflections and localized high-intensity echo responses, which may produce large local residual components in latent velocity regression. Under this condition, the conventional squared L_2 loss may assign disproportionate weights to these localized residuals, causing the optimization to be overly influenced by isolated strong response. Recent studies on improved rectified flow training have shown that the Pseudo-Huber loss provides a robust alternative to the squared L_2 objective by reducing the influence of large

residuals while retaining smooth quadratic behavior around small errors (Lee et al., 2024; Song et al., 2023). Inspired by this design, this study adopts an element-wise Pseudo-Huber loss in the latent velocity space:

$$L_{PH}(\delta) = \delta^2\left(\sqrt{1+\left(\frac{r}{\delta}\right)^2}-1\right) \tag{7}$$

where $r = v_\theta(z_t, t) - (\varepsilon - z)$ is the velocity-regression residual, and δ controls the transition between the quadratic and approximately linear regimes. In implementation, Eq. (7) is applied to each residual element and then reduced in the same manner as the other loss functions. Unless otherwise specified, δ is fixed to 1×10^{-4}. As indicated by Eq. (7), the loss behaves quadratically for small residuals and approximately linearly for large residuals. For GPR B-scan image generation, this property reduces the influence of isolated strong reflections and clutter-induced responses, encouraging the model to focus on the consistent reconstruction of defect-related structural patterns.

2.2.4. U-shaped timestep sampling for boundary emphasis

The distribution of training timesteps plays a critical role in rectified flow training. Although uniform sampling is commonly adopted, recent studies on improved rectified flow training have shown that boundary timesteps are important for few-step generation. To emphasize both the data and noise boundaries (Lee et al., 2024), this study adopts a symmetric Beta distribution for timestep sampling:

$$t \sim Beta(\alpha, \alpha), 0 < \alpha < 1 \tag{8}$$

where α controls the shape of the timestep distribution. When $\alpha = 1$, the distribution degenerates to uniform sampling. When $\alpha < 1$, more samples are assigned near the two boundaries ($t \to 0$ and $t \to 1$). Near the noise boundary ($t \to 1$), the model learns robust representations of speckle noise and clutter statistics. Near the data boundary ($t \to 0$), the model focuses on refining defect-related echo structures. Fig. 4 provides a schematic illustration of this U-shaped sampling strategy.

2.2.5. ODE-based generation with dual-text-encoder conditioning

During inference, new GPR B-scan samples are generated by numerically integrating the learned velocity field from an initial noisy latent state z_1 toward the clean latent manifold z_0. Following the rectified flow formulation in Section 2.2.1, the latent generation trajectory is governed by:

$$\frac{dz_t}{dt} = v_\theta(z_t, t, c) \tag{9}$$

To solve this ODE, this study adopts Heun's method, a classical second-order explicit Runge-Kutta scheme. Karras et al. introduced this method into diffusion-based generative ODE sampling and showed that it provides a favorable tradeoff between truncation error and the number of function evaluations (NFE), making it suitable to few-step generation (Karras et al., 2022). In intuitive terms, Heun's method first predicts the next state with a Euler step and then corrects the update using the average velocity over the interval. This property is particularly relevant to GPR-ImaGen, where efficient few-step latent sampling is preferred for practical GPR B-scan generation.

During inference, the time schedule follows the convention in Eq. (3), where $t = 1$ corresponds to the noisy latent state and $t = 0$ corresponds to the clean latent representation. Therefore, generation is performed along the backward sampling direction of the rectified flow ODE, from the noise boundary to the data boundary (Liu et al., 2023). Specifically, let t_k and t_{k+1} denote two adjacent sampling times with $t_k > t_{k+1}$, and define the backward integration step as $h_k = t_{k+1} - t_k < 0$. A predictor state is first computed by:

$$\tilde{z}_{t_{k+1}} = z_{t_k} + h_k v_\theta(z_{t_k}, t_k, c) \tag{10}$$

and the corrected update is then given by:

$$z_{t_{k+1}} = z_{t_k} + \frac{h_k}{2}\left[v_\theta(z_{t_k}, t_k, c) + v_\theta(\tilde{z}_{t_{k+1}}, t_{k+1}, c)\right] \tag{11}$$

After the latent trajectory reaches $t = 0$, the resulting latent code is decoded by the VAE decoder to obtain the generated GPR B-scan images (Rombach et al., 2022).

In the Flux.1-dev architecture, generation is further guided by structured textual prompts through a dual-text-encoder conditioning mechanism. Specifically, CLIP provides a pooled global text representation, while T5 provides sequence-level textual embeddings, forming complementary conditioning pathways for the rectified flow transformer. The pooled CLIP representation is used as a global conditioning vector together with timestep and guidance embeddings, while the T5 sequence embeddings are projected as text tokens and interact with image latent tokens through the double-stream and single-stream transformer blocks. For the GPR B-scan generation task, prompts describing defect category, approximate location, and spatial scale therefore serve as inspection-relevant conditioning signals. This

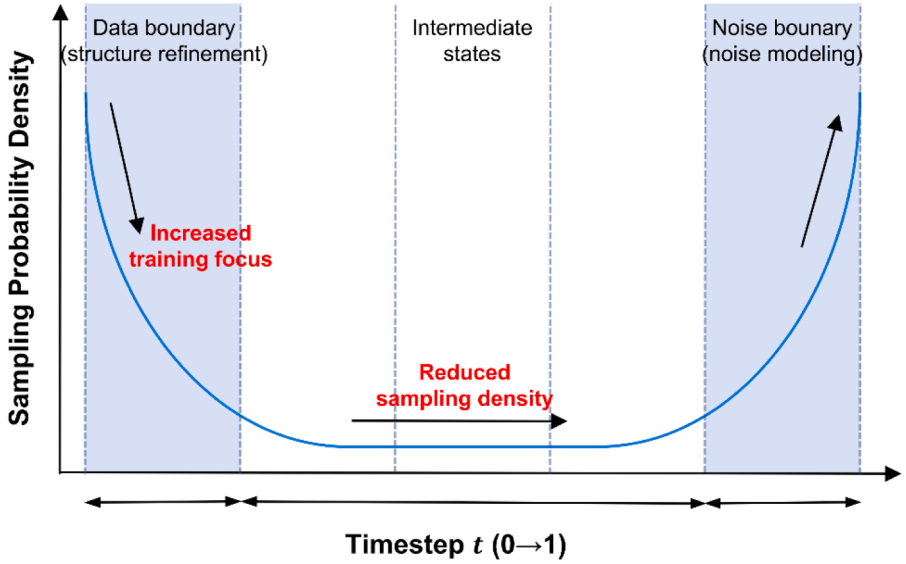


Fig. 4. Schematic illustration of the U-shaped timestep sampling strategy.

conditioning mechanism guides the learned velocity field during the ODE integration process in Eq. (10) and Eq. (11), enabling category-aware GPR B-scan image generation. Fig. 5 qualitatively demonstrates this text-guided generation behavior, where prompt-level semantic changes lead to corresponding variations in the generated radar structures.

2.2.6. Training configuration and computational efficiency

In this study, all experiments were conducted under a fixed training configuration. Specifically, all GPR B-scan images were used at their original resolution of 224 × ,224. DoRA was configured with a rank of 4 and was mainly applied to the linear projection layers in the double-stream and single-stream blocks of the Flux.1-dev backbone, as well as the attention and MLP modules of the text encoders. The model was optimized with AdamW 8bit using a learning rate of 8×10^{-4} and a cosine scheduler. The total number of training epochs was set to 32. During generation, 12 sampling steps and a guidance scale of 6.0 were adopted. Using the original B-scan resolution avoids additional resizing-induced distortion of radar reflection structures, while the relatively strong guidance scale reinforces the engineering prompt condition and constrains the sampling process toward category-consistent radargram structures. These settings were kept unchanged across experiments to improve reproducibility and enable a fair comparison of the proposed training components.

To support the selection of the adopted few-step generation setting, inference efficiency was measured by varying only the number of sampling steps within the same Flux.1-dev backbone. Specifically, the 12-step setting used in GPR-ImaGen was compared with a 50-step setting under the same prompt, image resolution, guidance scale, batch size, and hardware environment. The number of velocity-network evaluation was recorded as the number of function evaluations (NFE), and the relative denoising FLOPs reduction was estimated from the NFE ratio because each evaluation uses the same transformer backbone at the same latent resolution. Compared with the 50-step setting, the 12-step setting reduced the average per-image generation time from 8.8 s/image to 2.3 s/image on an RTX 4090 24 GB GPU, corresponding to a relative denoising FLOPs reduction of approximately 76%. The peak inference memory was approximately 21.7 GB, and the DoRA fine-tuning process required approximately 4.5 h with a peak GPU memory of 19.2 GB. These settings were kept unchanged across experiments to improve reproducibility and provide a practical configuration for efficient GPR B-scan generation.

2.3. Validation framework for GPR-ImaGen

2.3.1. Degradation-aware B-scan-adapted distribution metrics

Distribution-level metrics are first adopted to evaluate whether the generated B-scan images are statistically aligned with real GPR data. Conventional Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) are computed using the ImageNet-pretrained InceptionV3 feature space. However, GPR B-scan images are grayscale radargrams dominated by hyperbolic reflections, layered interfaces, directional echo textures, clutter, and speckle-like noise, which are substantially different from natural-image semantics. Therefore, directly relying on ImageNet-based features may not fully reflect the radar-specific structural consistency of generated B-scan images. To provide a complementary domain-adapted assessment, this study further constructs Bscan-FID and Bscan-KID using a degradation-aware B-scan-adapted InceptionV3 feature extractor.

Specifically, an ImageNet-pretrained InceptionV3 is fine-tuned on

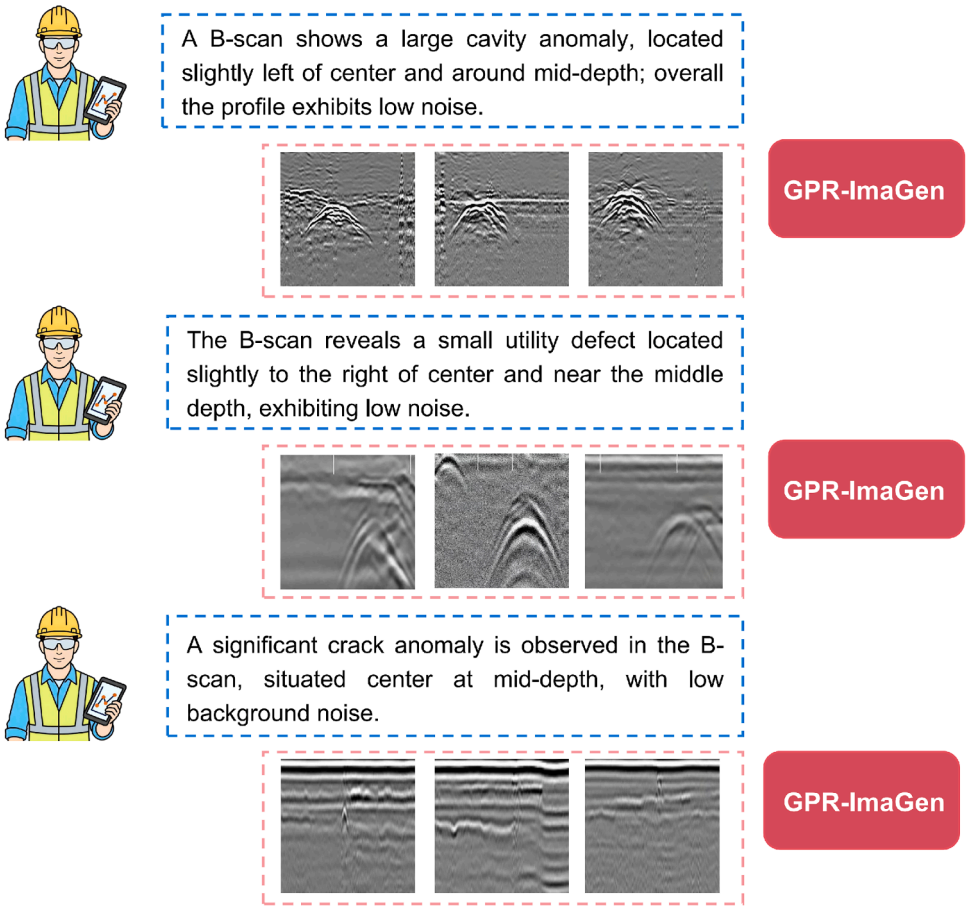


Fig. 5. Examples of text-guided controllable GPR B-scan generation.

real B-scan images through a controlled degradation-aware self-supervised task, where automatically generated degradation labels provide the supervision for learning radargram-specific feature representations. For a real B-scan image x, a degradation operation $A_{d,q}(\cdot)$ is applied to generate a degraded sample:

$$\widetilde{x} = A_{d,q}(x) \tag{12}$$

where d denotes the degradation type and q denotes the degradation level. The degradation types include directional blur, layer smoothing, local echo attenuation, Gaussian noise, and contrast shift. The InceptionV3 backbone $f_B(\cdot)$ extracts the feature representation, and two prediction heads are used to classify the degradation type and degradation level. The training objective is formulated as:

$$L_{DA} = \lambda_d L_{CE}(h_d(f_B(\widetilde{x})), d) + \lambda_q L_{CE}\left(h_q(f_B(\widetilde{x})), q\right) \tag{13}$$

where $h_d(\cdot)$ and $h_q(\cdot)$ are the degradation-type and degradation-level prediction heads, respectively; λ_d and λ_q are the corresponding loss weights. After training, only the InceptionV3 backbone is retained as the B-scan-adapted feature extractor, while the two auxiliary prediction heads are discarded.

Let R and G denote the real and generated B-scan image sets, respectively. The ImageNet-pretrained and B-scan-adapted InceptionV3 feature extractors are denoted as $\varphi_I(\cdot)$ and $\varphi_B(\cdot)$. The mean vectors and covariance matrices of real and generated features in the corresponding feature space are denoted as μ_r, μ_g, $\sum_r$, and $\sum_g$. The conventional FID is computed in the ImageNet feature space as (Heusel et al., 2017):

$$FID = \| \mu_r^I - \mu_g^I \|_2^2 + T_r\left(\sum_r^I + \sum_g^I - 2\left(\sum_r^I \sum_g^I \right)^{\frac{1}{2}} \right) \tag{14}$$

The conventional KID is computed in the same ImageNet feature space as (Bińkowski et al., 2018):

$$KID = \frac{1}{S} \sum_{s=1}^{S} MMD_u^2\left(Z_{r,s}^I, Z_{g,s}^I; k\right) \tag{15}$$

where $Z_{r,s}^I$ and $Z_{g,s}^I$ denote randomly sampled real and generated feature subsets extracted by $\varphi_I(\cdot)$, S is the number of random subsets, and k is the polynomial kernel.

Similarly, Bscan-FID is computed by replacing the ImageNet feature extractor with the degradation-aware B-scan-adapted feature extractor:

$$Bscan-FID = \| \mu_r^B - \mu_g^B \|_2^2 + T_r\left(\sum_r^B + \sum_g^B - 2\left(\sum_r^B \sum_g^B \right)^{\frac{1}{2}} \right) \tag{16}$$

Bscan-KID is defined as:

$$Bscan-KID = \frac{1}{S} \sum_{s=1}^{S} MMD_u^2\left(Z_{r,s}^B, Z_{g,s}^B; k\right) \tag{17}$$

where the superscripts I and B indicate the ImageNet-pretrained feature space and the B-scan-adapted feature space, respectively. Therefore, FID and KID measure distribution alignment from the perspective of general visual representations, whereas Bscan-FID and Bscan-KID provide a complementary assessment of radargram-specific structural consistency. These four metrics are jointly used to evaluate whether the generated B-scan images preserve both general visual quality and domain-relevant GPR characteristics.

2.3.2. Multi-level validation protocol

To comprehensively assess the practical value of the proposed generative framework, a multi-level validation protocol is established. Rather than relying on a single metric or a single experimental setting, the protocol is designed to examine the proposed method from two complementary perspectives: generation quality and downstream augmentation utility (Dai et al., 2025; Khosravi et al., 2025). Accordingly, the evaluation progresses from distribution-level assessment of generated B-scan images to application-oriented validation in defect detection under varying data availability, class-imbalance, and cross-domain conditions.

(1) Component ablation and hyperparameter sensitivity analysis. The first level evaluates the generation model, aiming to verify whether the proposed adaptation strategy enhances the domain adaptation capability of the pretrained Flux.1-dev model for GPR B-scan images. This level focuses on distribution alignment and the preservation of radar-specific structural characteristics. The main training components, including DoRA, Pseudo-Huber loss, and U-shaped timestep sampling, are ablated under the same data and generation settings. In addition, a lightweight sensitivity analysis is conducted for two key hyperparameters: the Pseudo-Huber transition threshold δ and the U-shaped timestep sampling parameter α. To isolate the influence of each hyperparameter, one parameter is varied at a time while the remaining training configuration is kept unchanged. Specifically, the effect of δ is evaluated under the default U-shaped timestep sampling setting $\alpha = 0.6$, whereas the effect of α is evaluated under the default Pseudo-Huber setting $\delta = 1 \times 10^{-4}$. The corresponding ablation and sensitivity results are reported in Section 3.1.2.

(2) Comparison with representative generative baselines. The second level compares the proposed method with representative generative models that have been used or widely considered for GPR-related data augmentation, including WGAN, DCGAN, and DDPM. All competing models are trained on the same preprocessed real GPR B-scan training data. WGAN is trained for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, $\beta_1 = 0.0$, $\beta_2 = 0.9$, and a latent dimension of 128. DCGAN is trained for 100 epochs using the Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.9$, a latent dimension of 100, and binary cross-entropy loss for adversarial learning. DDPM is implemented with a U-Net-based noise-prediction network and trained using AdamW for 100,000 optimization steps. The diffusion process is discretized into 1000 timesteps, and an exponential average (EMA) of the model parameters is maintained during training. The generated samples are evaluated using FID, KID, Bscan-FID, and Bscan-KID. In addition, t-distributed stochastic neighbor embedding (t-SNE) is used as an auxiliary visualization tool to project high-dimensional deep features into a two-dimensional space, so that the distribution coverage and local neighborhood relationship between real and generated samples can be intuitively examined. The results of this level are reported in Section 3.1.3.

(3) Memorization and intra-prompt diversity assessment. The third level evaluates the reliability of generated B-scan images from the perspective of memorization and intra-prompt diversity. Since the proposed framework is fine-tuned using a limited number of real GPR B-scan images, it is necessary to examine whether the generated samples are high-similarity copies of the training images. For this purpose, structural similarity between two normalized B-scan images is first measured using SSIM (Wang et al., 2004):

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{18}$$

where x and y denote two normalized B-scan images; μ_x and μ_y are their mean intensities; σ_x^2 and σ_y^2 are their variances; σ_{xy} is their covariance; C_1 and C_2 are small constants used for numerical stability.

Let G denote the generated image set and R denote the real training set. For each generated image, its maximum SSIM value with all real training images is calculated as the generated-to-real nearest-neighbor similarity. Meanwhile, real-to-real nearest-neighbor SSIM is computed within the real training set as a reference for the intrinsic similarity level of real GPR B-scan data:

$$S_i^{G\to R} = \max_{1\le j\le N_r} SSIM(g_i, r_j) \tag{19}$$

$$S_i^{R\to R} = \max_{\substack{1\le j\le N_r,\\ j\neq i}} SSIM(r_i, r_j) \tag{20}$$

where g_i, r_i and r_j denote generated and real images, respectively. A lower generated-to-real nearest-neighbor similarity than the real-to-real reference indicates that the generated samples are unlikely to be direct copies of the training images.

In addition, to memorization assessment, intra-prompt diversity is evaluated using multi-scale structural similarity (MS-SSIM) (Wang et al., 2003). MS-SSIM extends SSIM to multiple image scales and has been used as a practical proxy for perceptual diversity in conditional image synthesis, where lower pairwise MS-SSIM indicates higher diversity among samples generated under the same condition (Odena et al., 2017). For a prompt p with n_p generated images produced using different random seeds, the mean pairwise MS-SSIM is calculated as:

$$\overline{M}_p = \frac{2}{n_p(n_p-1)} \sum_{1\le i<j\le n_p} MS-SSIM\left(x_i^{(p)}, x_j^{(p)}\right) \tag{21}$$

where $x_i^{(p)}$ and $x_j^{(p)}$ denote two generated images conditioned on the same prompt. A lower $\overline{M}_p$ indicates stronger intra-prompt diversity, whereas an excessively high value suggests that the model may generate highly similar or collapsed outputs under the same textual condition. In this study, pairwise similarity values higher than 0.95 are regarded as high-similarity duplicate candidates. The results of memorization and intra-prompt diversity assessment are reported in Section 3.1.4.

(4) Downstream augmentation across multiple detection models. The fourth level shifts the focus from the quality of generated images to its practical function as a data augmentation tool for downstream detection. Generated GPR B-scan images are introduced into four representative object detection frameworks, including YOLOv8, YOLOv10, YOLOv11, and Faster R-CNN. These models represent different detection paradigms, enabling a comprehensive evaluation of whether the proposed augmentation strategy generalizes across detector architectures. The original dataset contains 873 B-scan images with six defect categories, among which 611 images are used for training. Synthetic samples are gradually added to the training set at ratios of approximately 25%, 50%, 75%, and 100% relative to the original training size. All generated B-scan images are manually annotated with defect categories and bounding boxes in YOLO format. The validation and test sets are kept unchanged and consist exclusively of real data to ensure fair evaluation. For consistency, YOLOv8, YOLOv10, and YOLOv11 are trained under the same settings, with 200 epochs, an input image size of 640 $\times$ 640, and a batch size of 64. Faster R-CNN is trained with a ResNet50 backbone for 150 epochs using an input image size of 640 $\times$ 640, and is optimized using Adam with an initial learning rate of 1 $\times 10^{-4}$ and a cosine learning-rate schedule.

Detection performance is evaluated using mAP@50, which measures the mean average precision at an intersection-over-union threshold of 0.50 (Lv et al., 2025). For category k, a predicted bounding box B_p is regarded as a valid detection when:

$$IoU(B_p, B_g) = \frac{|B_p \cap B_g|}{B_p \cup B_g} \ge 0.50 \tag{22}$$

where B_g denotes the corresponding ground-truth bounding box. The category-wise average precision is calculated as the area under the precision-recall curve (Lv et al., 2025):

$$AP_k = \int_0^1 P_k(R)dR \tag{23}$$

and mAP@50 is obtained by averaging the AP values over all K defect categories (Lv et al., 2025):

$$mAP@50 = \frac{1}{K}\sum_{k=1}^{K} AP_k|_{IoU=0.50} \tag{24}$$

This metric is used as the main task-oriented indicator for evaluating whether generated B-scan images improve defect recognition and localization on real test data. The corresponding results are reported in Section 3.2.1.

(5) Two-dimensional evaluation under data scarcity conditions. The fifth level further examines how the effectiveness of synthetic augmentation varies with the amount of available real data. A two-dimensional evaluation setting is constructed by jointly varying real-data retention and generated-data supplementation. This design makes it possible to distinguish two complementary roles of generative augmentation: compensating for missing training information when real data are scarce, and expanding distributional diversity when real data are relatively sufficient. Specifically, four real-data retention levels are constructed by retaining 100%, 75%, 50%, and 25% of the original training set. Let $N = 611$ denote the original real training set size. For each real-data retention ratio r, the "Origin" setting contains rN real images and no generated images. The "Mix" setting contains rN real images and $(1-r)N$ generated images, resulting in a total training size of N, respectively. The "Mix+25%," "Mix+50%," "Mix+75%," and "Mix+100%" settings further introduce additional generated images corresponding to 25%, 50%, 75%, and 100% of N. The corresponding results are reported in Section 3.2.2.

(6) Category-specific augmentation under class-imbalance conditions. The sixth level addresses category-specific augmentation under class-imbalance conditions. Instead of treating data deficiency only as a global problem, this level evaluates whether the proposed text-conditioned generation mechanism can provide semantically consistent synthetic samples for targeted defect categories. For each defect category, the corresponding class is reduced to 25% of its original training size, while all other categories and the validation and test sets are kept unchanged. Based on this setting, category-consistent synthetic augmentation is introduced by progressively adding generated B-scan images for the data-scarce class at ratios of 50%, 100%, 150%, 200%, 300%, and 400% relative to the remaining real samples. All synthetic samples are randomly selected from the corresponding category, and detection performance is evaluated using mAP@50. This level aims to reveal whether category-aware GPR B-scan image generation can effectively mitigate class-imbalance and support the learning of minority-class radar signatures. The results are reported in Section 3.2.3.

(7) Cross-domain generalization under domain shift. The seventh level evaluates cross-domain generalization under domain shift. This level first quantifies the feature-distribution discrepancy between the source and target domains before conducting cross-domain detection experiments. Specifically, FID and Bscan-FID are computed among three distribution pairs: source-domain split A versus source-domain split B, source-domain real images versus target-domain real images, and generated images versus target-domain real images. The intra-source distance provides a reference for natural variation within the source domain, while the source-target and synthetic-target distances characterize the domain gap faced by cross-domain augmentation. Based on the quantified domain gap, cross-domain detection evaluation is conducted on a held-out target-domain GPR B-scan dataset. This target-domain dataset contains 500 real B-scan images from three categories, including crack, interlayer debonding, and utility. Although these samples are selected from the same two public datasets, they are strictly separated from the images used for GPR-ImaGen fine-tuning and are reserved only for cross-domain detection evaluation. Synthetic samples from the corresponding categories are progressively introduced into the target-domain training set at 25%, 50%, 75%, and 100% augmentation ratios, whereas the validation and test sets remain composed only of real target-domain images. This level therefore examines whether generated data can provide transferable defect-related structural information

under domain shift, and whether excessive synthetic augmentation may introduce source-domain bias. The results are reported in Section 3.2.4.

3. Results and discussion

3.1. Image-quality evaluation of generated B-scan data

3.1.1. Validation of the B-scan-adapted inceptionV3 feature extractor

This section presents the validation results of the degradation-aware B-scan-adapted InceptionV3 feature extractor described in Section 2.3.1. As shown in Fig. 6(a), both training and validation accuracy increase rapidly during the early training stage and then gradually remains mostly within the range of 0.85–0.90, while the training accuracy follows a similar trend. The close agreement between the two curves indicates that the model does not simply memorize the training samples, but learns transferable degradation-sensitive representations from real B-scan images. This result suggests that the adapted feature extractor can capture domain-relevant variations associated with B-scan image degradation.

The confusion matrix of the best validation model is shown in Fig. 6 (b). The matrix exhibits clear diagonal dominance, indicating that most degraded samples are assigned to their corresponding degradation types. In particular, directional blur and local echo attenuation are correctly classified, indicating that the adapted feature extractor is sensitive to directional texture degradation and localized echo-energy variations. Gaussian noise, original images, and layer smoothing are also well recognized, suggesting that the model can distinguish changes in radargram noise distribution, layered structures, and unmodified B-scan patterns. Although contrast shift shows relatively more confusion with several other classes, this is reasonable because real B-scan images naturally contain gray-level variations caused by acquisition conditions, gain adjustment, and preprocessing. Overall, these results demonstrate that the fine-tuned InceptionV3 backbone acquires discriminative sensitivity to radargram-specific degradation patterns. Therefore, the resulting B-scan-adapted feature space provides a more domain-relevant basis for Bscan-FID and Bscan-KID, complementing conventional FID and KID in the subsequent evaluation of generated B-scan images.

3.1.2. Component ablation and hyperparameter sensitivity analysis

After validating the degradation-aware B-scan-adapted feature extractor, the contribution of each training component is further examined using both conventional distribution metrics and B-scan-adapted distribution metrics. Table 2 summarizes the ablation results of different training configurations under FID, KID, Bscan-FID, and Bscan-KID. The combination of DoRA, Pseudo-Huber loss, and U-shaped timestep sampling achieves the strongest overall performance, with an FID of 10.4958, KID of 0.0026, Bscan-FID of 10.8279, and Bscan-KID of 0.0149. Although DoRA+L2+U-shaped obtains a slightly lower Bscan-FID, the full configuration achieves the best FID, KID, and Bscan-KID, while maintaining highly competitive Bscan-FID. This indicates that the proposed training framework provides a more balanced improvement in both general visual distribution alignment and radargram-specific feature consistency. In particular, the reduction in Bscan-KID suggests that the generated samples better preserve radargram-specific structural characteristics, including layered textures, local echo responses, and background clutter patterns.

Table 2
Ablation study results.

Experiment	FID (↓)	KID (↓)	Bscan-FID (↓)	Bscan-KID (↓)
LoRA+L2+Shift	13.3716	0.0089	13.7205	0.0223
DoRA+L2+Shift	13.4413	0.0085	16.6200	0.0293
DoRA+L2+U-shaped	13.4682	0.0060	**10.7267**	0.0151
DoRA+Pseudo-Huber+Shift	11.1349	0.0045	12.8080	0.0201
LoRA+Pseudo-Huber+U-shaped	14.3329	0.0064	13.8938	0.0226
DoRA+Pseudo-Huber+U-shaped	**10.4958**	**0.0026**	10.8279	**0.0149**

The sensitivity analysis of the Pseudo-Huber transition threshold δ and the U-shaped timestep parameter α is summarized in Table 3. For the Pseudo-Huber loss, $\delta = 1 \times 10^{-4}$ achieves the lowest FID and Bscan-FID among the tested settings, indicating that it provides an appropriate balance between residual robustness and structural sensitivity. A smaller δ imposes stronger suppression on large residuals, which may weaken the learning of moderate defect-related structures, whereas a larger δ makes the loss closer to the squared L_2 loss and increases sensitivity to local high-intensity responses. For the timestep distribution, $\alpha = 0.6$ gives the best overall distribution consistency. Compared with $\alpha = 0.2$

Table 3
Sensitivity analysis of δ and α.

Analysis group	δ	α	FID	Bscan-FID
δ sensitivity	1×10^{-5}	0.6	10.6118	13.2941
δ sensitivity	1×10^{-3}	0.6	11.5056	12.6674
α sensitivity	1×10^{-4}	0.2	11.2756	11.1280
α sensitivity	1×10^{-4}	0.4	13.4327	14.0090
α sensitivity	1×10^{-4}	0.8	13.2031	11.3487
Default setting	1×10^{-4}	0.6	**10.4958**	**10.8279**

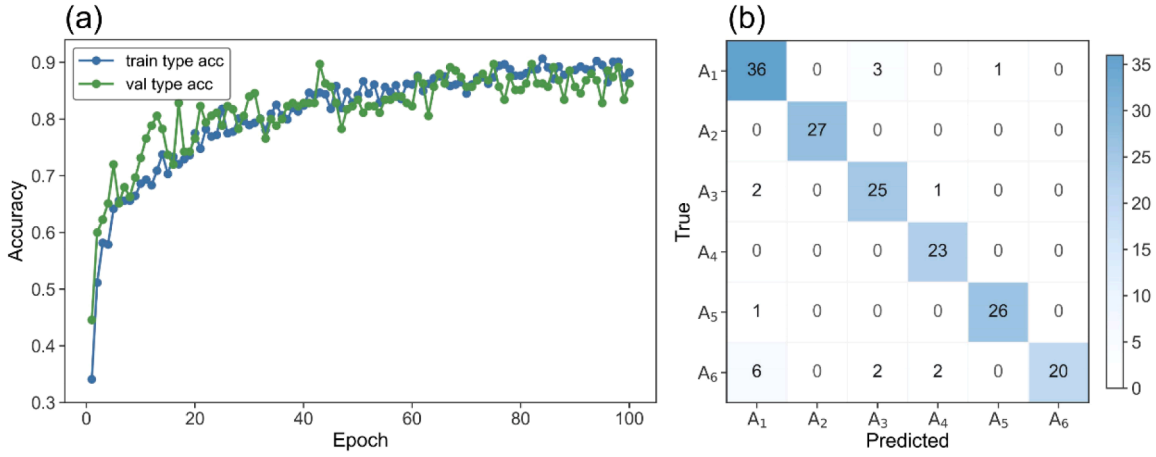


Fig. 6. Validation of the degradation-aware B-scan-adapted InceptionV3 feature extractor. (a) Degradation-type prediction accuracy during self-supervised fine-tuning. (b) Confusion matrix of degradation-type classification on the validation set. $A_1 - A_6$ denote original, directional blur, layer smoothing, local echo attenuation, Gaussian noise, and contrast shift respectively.

and $\alpha = 0.4$, it avoids excessive concentration of training samples near the two boundaries. Compared with $\alpha = 0.8$, it provides stronger boundary emphasis for few-step generation. These results support the use of $\delta = 1 \times 10^{-4}$ and $\alpha = 0.6$ as the default settings in GPR-ImaGen.

3.1.3. Comparative analysis with baseline generative models

Following the baseline comparison protocol described in Section 2.3.2, this section compares GPR-ImaGen with representative generative models using both distribution metrics and feature-space visualization. The quantitative results are summarized in Table 4. It can be observed that GPR-ImaGen achieves the best performance among all compared models, with FID of 10.4958, KID of 0.0026, Bscan-FID of 10.8279, and Bscan-KID of 0.0149. Compared with WGAN and DCGAN, the proposed method shows a clear reduction in both conventional and B-scan-adapted distribution metrics, indicating better modeling of heterogeneous radargram structures when only 873 real B-scan images are available for generative-model training. DDPM achieves better distribution alignment than the adversarial baselines, but its FID and Bscan-FID are still higher than those of GPR-ImaGen. Specifically, GPR-ImaGen reduces FID from 18.4147 to 10.4958 and Bscan-FID from 21.5629 to 10.8279 compared with DDPM. The larger improvement in the B-scan-adapted feature space further suggests that the generated samples better preserve radargram-specific structures, such as layered interfaces, local echo responses, and speckle-like background variations. These results demonstrate the advantage of the proposed category-aware rectified flow framework for limited-data GPR B-scan generation.

The t-SNE visualization in Fig. 7 provides an intuitive comparison of feature-space distribution coverage. The real B-scan samples form a multi-cluster distribution in the embedding space, where different clusters correspond to different defect categories. This pattern reflects the inherent inter-class variation of GPR B-scan images in terms of echo morphology, texture characteristics, and local structural responses. Against this reference distribution, the samples generated by the proposed method are more broadly distributed around multiple real-data clusters and exhibit comparatively balanced coverage across feature regions dominated by different defect categories. In contrast, WGAN and DCGAN tend to concentrate in limited local regions, failing to cover the full data manifold, while DDPM shows partial alignment but remains relatively constrained. This broader and more balanced embedding distribution suggests that the proposed method achieves stronger mode coverage and better preserves category-relevant structural features, rather than merely reproducing generic radar textures or averaged echo pattern. Therefore, Fig. 7 provides feature-distribution-level evidence that the proposed framework offers clear advantages in diversity, realism, and category-sensitive structural representation, which also helps explain its superior effectiveness in the subsequent data augmentation experiments for downstream defect detection.

3.1.4. Memorization and intra-prompt diversity assessment

Following the SSIM-based reliability assessment described in Section 2.3.2, this section reports the nearest-neighbor memorization analysis and intra-prompt diversity results. Table 5 presents the nearest-neighbor SSIM results for memorization assessment. The generated-to-real comparison gives a mean SSIM of 0.321 and a maximum SSIM of 0.658, and no generated sample exceeds the high-similarity threshold of 0.95. In contrast, the real-to-real reference comparison shows a higher mean SSIM of 0.516 and a maximum SSIM of 0.998, with 12 high-similarity pairs observed within the real dataset. This indicates that the generated samples do not show stronger similarity to the training set than the intrinsic similarity already existing among real B-scan images, suggesting that GPR-ImaGen does not simply memorize the training data.

Table 6 reports the intra-prompt diversity results on the pairwise MS-SSIM. The mean pairwise MS-SSIM values range from 0.046 to 0.176 across the six defect categories, and no image pair exceeds the high-similarity threshold of 0.95. These results indicate that GPR-ImaGen does not collapse to nearly identical outputs under the same prompt and can generate structurally diverse B-scan images under identical textual conditions. Among the categories, loose, cavity, and utility show relatively lower MS-SSIM values, suggesting stronger intra-prompt variation, whereas crack, void, and interlayer debonding exhibit slightly higher similarity, likely due to more constrained radar response patterns or more consistent prompt-level structural descriptions. Overall, the generated samples provide additional structural variations beyond simple replication while maintaining category-level semantic consistency.

Table 4
Comparison of generative models.

Model	FID	KID	Bscan-FID	Bscan-KID
WGAN	26.9333	0.0273	54.6412	0.1124
DCGAN	21.6904	0.0168	35.1113	0.0726
DDPM	18.4147	0.0159	21.5629	0.0407
Ours	**10.4958**	**0.0026**	**10.8279**	**0.0149**

3.2. Validation of generated data augmentation

3.2.1. Augmentation performance across multiple detection models

Following the multi-model augmentation protocol described in Section 2.3.2, this section reports the effect of generated B-scan images on downstream defect detection. Since the validation and test sets consist only of real B-scan images, the changes in mAP@50 directly reflect whether the generated samples can improve defect recognition and localization in real GPR data. The detection results are presented in Fig. 8. It can be observed that incorporating generated B-scan images improves the detection performance of all four detector architectures at certain augmentation ratios. Among them, YOLOv11 achieves the highest detection performance, with an mAP@50 of 91.6% at the 100% augmentation ratio. These improvements indicate that the generated samples provide useful defect-related reflection patterns and complementary structural variations, thereby enriching the training distribution and enhancing the learning of radar responses associated with subsurface defects. However, the performance gain is not strictly proportional to the augmentation ratio. When the amount of generated data gradually approaches the scale of real training data, the improvement tends to saturate, suggesting that synthetic B-scan images can effectively supplement the training set, but real radar observations remain essential for preserving domain-specific signal characteristics, background clutter, and acquisition-related variations.

Overall, the consistent performance improvement observed in both YOLO-family one-stage detectors and the two-stage Faster R-CNN detector demonstrates that the proposed augmentation strategy is not restricted to a specific detection architecture. These results support the practical utility of GPR-ImaGen as a detection-oriented data augmentation tool for intelligent GPR interpretation.

3.2.2. Two-dimensional evaluation under data scarcity conditions

Following the two-dimensional data-scarcity evaluation protocol defined in Section 2.3.2, this section reports the detection performance under different real-data retention levels and synthetic augmentation settings. The experimental results are summarized in Fig. 9 and two key observations are identified. When real training data are limited, incorporating generated B-scan images leads to substantial improvements in detection accuracy. The improvement is particularly pronounced in the 25% and 50% real-data scenarios, indicating that synthetic samples effectively compensate for insufficient training samples and structural diversity. Furthermore, when more real data are retained, synthetic data still contributes to performance improvements by enriching the diversity of radar reflection structures. However, the marginal benefit decreases as the amount of real data increases. This indicates that generated samples mainly serve to expand the training distribution rather than directly replace real radar observations. Notably, under the

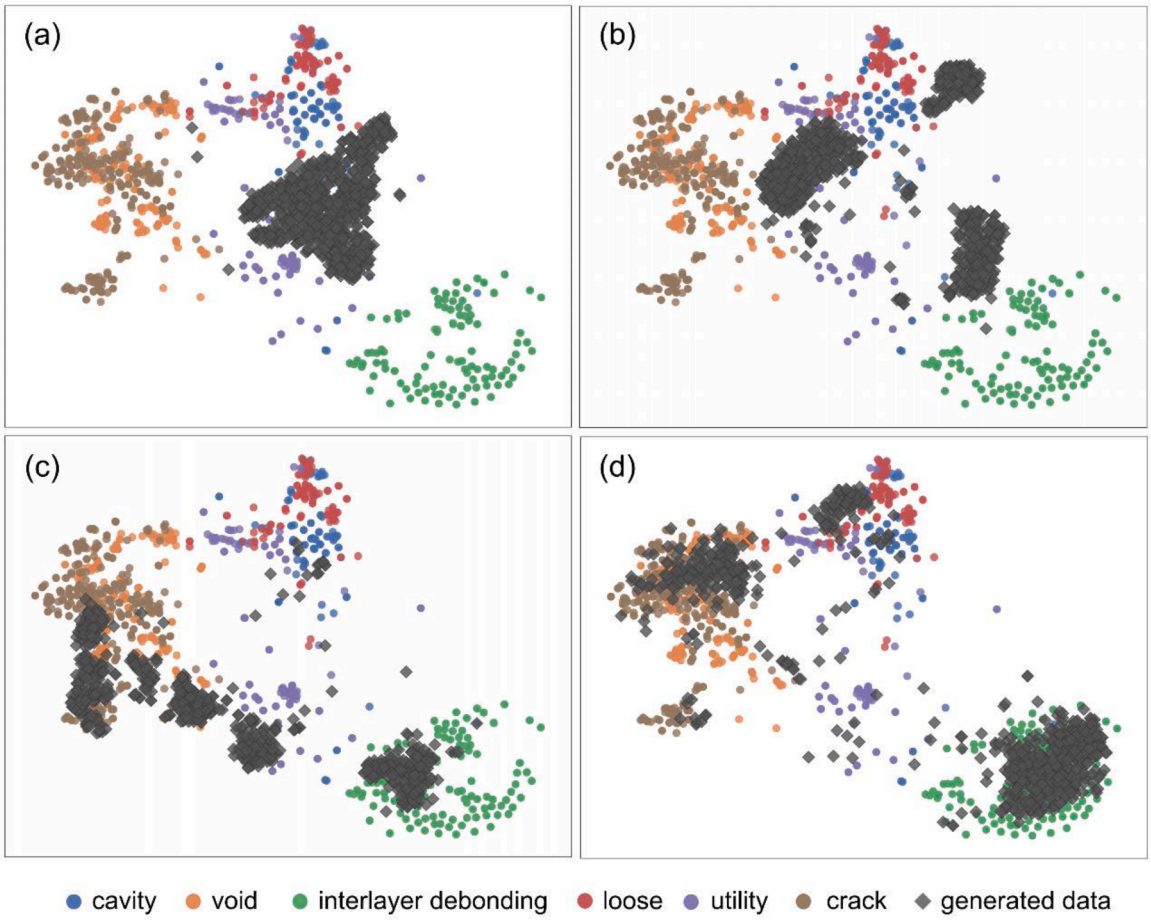


Fig. 7. Two-dimensional t-SNE visualization of real and synthetic GPR data by different models. (a) WGAN. (b) DCGAN. (c) DDPM. (d) Ours.

Table 5
Nearest-neighbor SSIM analysis for memorization assessment.

Comparison	Mean SSIM	Max SSIM	SSIM>0.95
$S^{G\to R}$	0.321	0.658	0
$S^{R\to R}$	0.516	0.998	12

Table 6
Pairwise SSIM for intra-prompt diversity assessment.

Category	Prompts/ Images	Mean Pairwise MS-SSIM	Max pairwise MS-SSIM	Pairs above 0.95
Cavity	5/50	0.085	0.218	0
Void	5/50	0.157	0.601	0
Interlayer debonding	5/50	0.154	0.632	0
Loose	5/50	0.046	0.167	0
Utility	5/50	0.090	0.409	0
Crack	5/50	0.176	0.612	0

75% real-data setting, the performance gain achieved by further synthetic augmentation is even greater than that obtained when augmentation is applied to the 100% real-data setting. A plausible explanation is that the retained training subset still preserves the main characteristics of real B-scan data after partial removal of real samples, while the generated data introduce a broader range of structural variations than the removed portion of real data. As a result, synthetic augmentation in this setting not only compensates for missing samples, but also effectively expands the coverage of the training distribution. This observation further supports the role of generated data as a mechanism for distribution expansion, especially when the real dataset remains sufficiently representative but no longer fully covers the underlying structural variability.

3.2.3. *Category-specific augmentation under class-imbalance*

Following the category-specific augmentation protocol described in Section 2.3.2, this section reports the effect of category-consistent generated samples under class-imbalance conditions. The resulting performance trends for all categories are summarized in Fig. 10. Under severe class reduction, detection accuracy drops markedly across all categories, confirming that the scarcity of class-specific samples directly impairs the learning of discriminative radar signatures. As category-consistent synthetic data are progressively introduced, performance generally recovers across all categories, indicating that the generated samples provide useful structural cues for compensating missing training information. However, the recovered performance usually remains below the original full-data baseline, suggesting that synthetic samples can alleviate class scarcity but cannot fully replace the diversity and fidelity of real observations.

The magnitude and rate of performance recovery further exhibit clear class-dependent variations. This reflects differences in radar signature characteristics and in the extent to which synthetic data can capture representative structural patterns. Categories with more distinctive geometric structures or stronger reflections tend to benefit from relatively low augmentation ratios, whereas categories dominated by diffuse, ambiguous, or noise-sensitive patterns generally require higher augmentation ratios to achieve comparable gains. Consequently, the effectiveness of synthetic augmentation does not follow a uniform

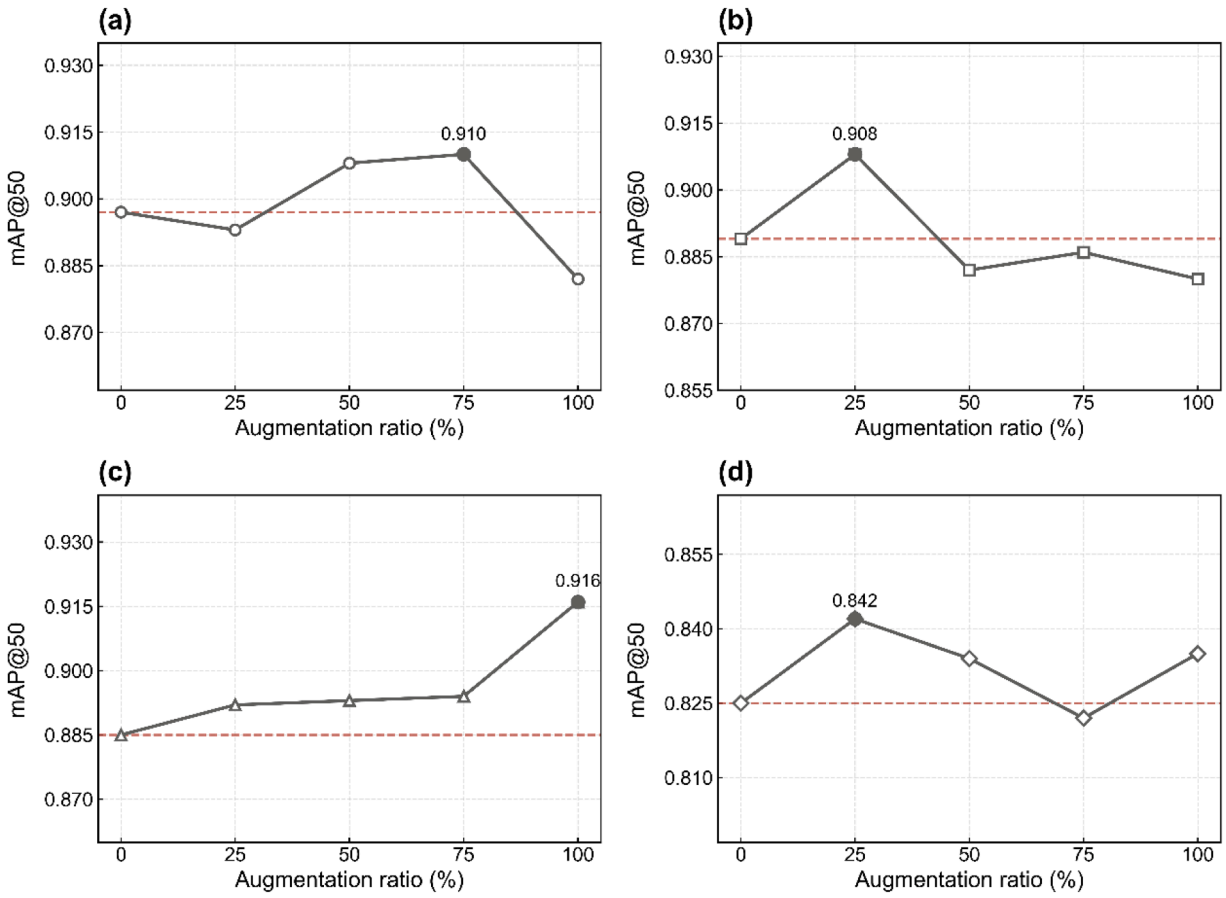


Fig. 8. Detection performance of four object detectors under different synthetic augmentation ratios. (a) YOLOv8. (b) YOLOv10. (c) YOLOv11. (d) Faster R-CNN.

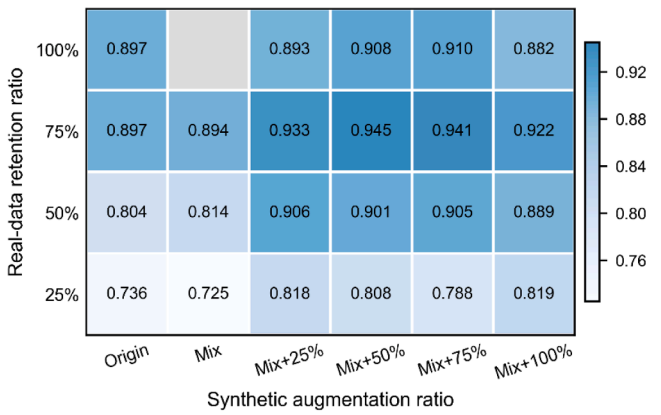


Fig. 9. Detection performance (mAP@50) under different real-data retention levels and synthetic augmentation settings. Blank cell indicates the setting is equivalent to the corresponding Origin baseline and is therefore omitted to avoid repetition.

trend, and the most beneficial augmentation ratio varies across defect categories. Despite these differences, all categories show an overall upward tendency with the introduction of synthetic data. This indicates that the generated samples capture transferable and category-relevant structural information rather than introducing spurious visual patterns. Overall, these results support the practical value of category-consistent generation for mitigating class imbalance in GPR defect detection.

3.2.4. *Cross-domain evaluation of synthetic data augmentation*

Following the cross-domain validation protocol defined in Section 2.3.2, this section reports the quantified source-target distribution discrepancy and the corresponding cross-domain detection results. Table 7 presents the quantitative characterization of the domain gap. The source-target distance is clearly larger than the intra-source reference distance. Specifically, FID increases from 13.522 for the intra-source split to 22.095 between source-domain and target-domain real images, while Bscan-FID increases from 10.133 to 13.485. This confirms that the target-domain data have a measurable distribution discrepancy from the source-domain data. The generated-to-target distance is further increased, with an FID of 31.714 and a Bscan-FID of 28.702, indicating that generated samples retain more source-domain characteristics and do not fully match the target-domain background and noise distribution.

These results indicate that the cross-domain experiment is conducted under a clear distribution shift, rather than a simple random split of the same data source. They also provide a basis for interpreting the subsequent detection results: generated samples may transfer defect-related structural information, but their background statistics and noise characteristics remain different from those of the target domain.

The cross-domain detection results are shown in Fig. 11. The target-domain dataset contains 500 real B-scan images from three categories, including crack, interlayer debonding, and utility. Without synthetic augmentation, the baseline model achieves an mAP@50 of 0.776. After introducing 25% and 50% synthetic data into the target-domain training set, mAP@50 increases to 0.814 and 0.832, respectively. This improvement shows that moderate synthetic augmentation can enhance cross-domain feature learning by introducing additional defect-related reflection patterns.

However, the performance decreases when the augmentation ratio is further increased. The mAP@50 drops to 0.804 at the 75%

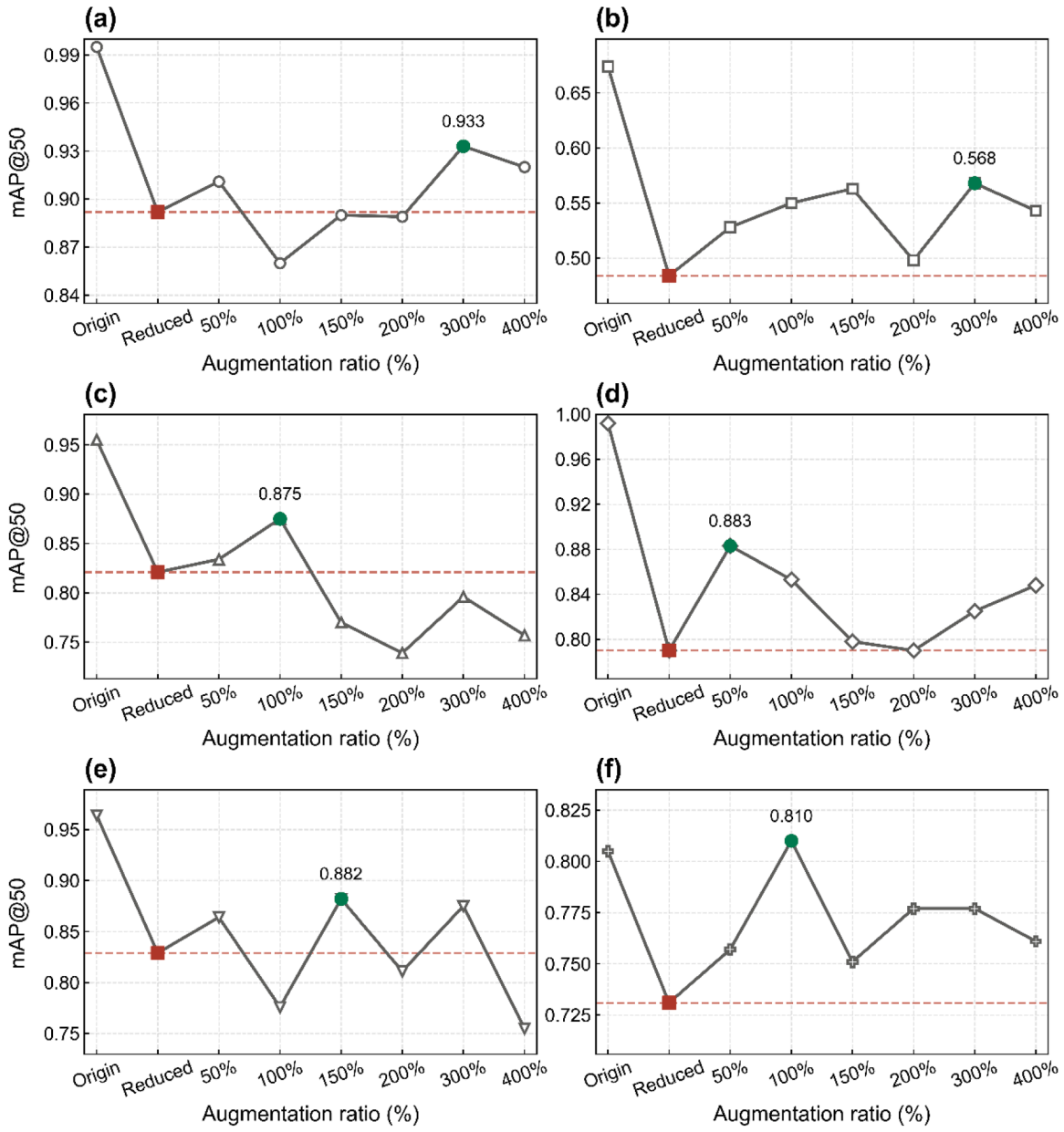


Fig. 10. Category-specific synthetic augmentation for mitigating class imbalance in GPR defect detection. (a) cavity. (b) void. (c) interlayer debonding. (d) loose. (e) utility. (f) crack.

Table 7
Quantitative characterization of domain gap using distribution distances.

Distribution pair	FID	Bscan-FID	Description
Source split A to B	13.5215±0.3377	10.1334±0.1621	Intra-source reference distance
Source real to Target real	22.0952	13.4846	Source-target domain discrepancy
Generated to Target real	31.7141	28.7021	Synthetic-target discrepancy

augmentation ratio and 0.778 at the 100% augmentation ratio. This non-monotonic trend is consistent with the domain-gap results in Table 7. At moderate ratios, generated samples act as complementary structural information and improve the learning of defect responses. At high ratios, the larger synthetic-target discrepancy may introduce excessive source-domain texture, noise, and reflection-style bias into the target-domain training set. As a result, the detector becomes less sensitive to authentic target-domain radar patterns. Overall, the cross-domain results show that GPR-ImaGen can improve detection under domain shift, but synthetic data should be used as a complementary source of structural diversity rather than as a dominant replacement for real target-domain observations.

4. Conclusions

Deep learning-based intelligent GPR detection has become an important direction for transportation infrastructure inspection, yet its practical deployment is still constrained by the scarcity and class imbalance of labeled GPR B-scan data. To address this challenge, this study proposed GPR-ImaGen, a rectified flow-based generative augmentation framework for GPR B-scan images. Specifically, the pre-trained text-to-image generative model (Flux.1-dev) was adapted to the GPR domain through DoRA-based parameter-efficient fine-tuning. To better accommodate the structural characteristics of GPR B-scan images and the training instability caused by limited data, a Pseudo-Huber loss

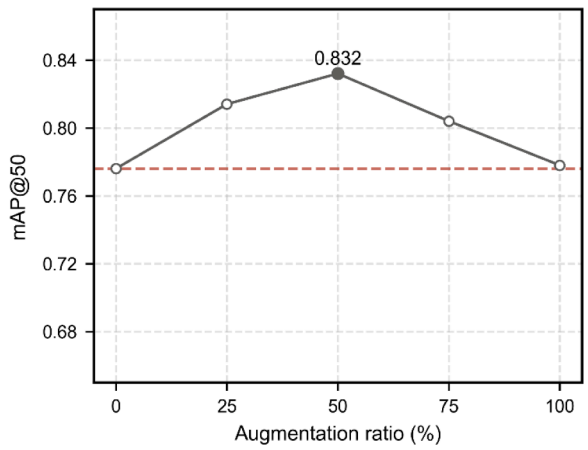


Fig. 11. Cross-domain detection performance under different generative augmentation ratios.

and a U-shaped timestep sampling strategy were further introduced as two key improvements to enhance training robustness and generation quality. The proposed framework achieved FID, KID, Bscan-FID, and Bscan-KID values of 10.4958, 0.0026, 10.8279, and 0.0149, showing the best overall distribution alignment among the compared generative models. In addition, the nearest-neighbor SSIM analysis and intra-prompt diversity assessment further indicated that the generated samples were not high-similarity copies of real training images and could provide structural variations under consistent textual conditions.

To verify the practical value of the proposed generative framework for downstream intelligent GPR detection, a series of augmentation experiments were conducted under different data availability and evaluation settings. By incorporating the generated B-scan images into training of detector, the proposed framework consistently improved detection performance across multiple detectors and augmentation ratios settings, with YOLOv11 achieving the highest mAP@50 of 91.6% at a 100% augmentation ratio. The improvement was particularly evident under data-scarce conditions, where mAP@50 increased from 73.6% to 81.9% when only 25% of the real training data were retained. Under class-imbalance settings, category-consistent augmentation also improved the detection of data-scarce categories, with mAP@50 increasing from 73.0% to 81.0% for crack. Cross-domain experiments further showed that generative data augmentation remained beneficial under domain shift, suggesting that the generated B-scan images capture more generalizable structural features rather than merely fitting the source-domain distribution. These results demonstrate that the proposed framework effectively complements limited real GPR data and provides practical support for improving both the detection accuracy and the generalization capability of intelligent GPR detection.

Overall, the proposed category-aware GPR B-scan image generation framework provides an effective and generalizable solution for alleviating data scarcity and class imbalance in deep learning-based intelligent GPR detection. Although the proposed framework has explored spatially aware generation under textual semantic guidance, its sensitivity to spatial and scale-related textual information remains limited. Future work will therefore incorporate richer GPR physical parameters and spatial-scale information to support more controllable generation in complex inspection scenarios. In addition, since the adopted open-source datasets mainly provide processed B-scan images rather than calibrated raw radar traces, this study focuses on radargram-level distribution consistency and detection-oriented utility, while more rigorous electromagnetic waveform-level validation will be investigated in future work when raw GPR acquisition records are available. With these improvements, the proposed framework is expected to further enhance the generalization and robustness of deep learning-based intelligent GPR detection in real engineering applications, and may also offer strong potential for other intelligent object detection tasks under data-scarce conditions.

CRediT authorship contribution statement

Zuhao Wang: Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Guangyao Li:** Writing – original draft, Validation, Investigation, Data curation, Conceptualization. **Fengzijun Pan:** Writing – original draft, Investigation, Formal analysis. **Zepeng Fan:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Xiaopei Cai:** Writing – review & editing, Validation, Supervision. **Dawei Wang:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Zepeng Fan reports financial support was provided by National Natural Science Foundation of China (U2569209). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (U2569209).

Data availability

Data will be made available on request.

References

Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying MMD GANs. *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lUOzWCW.

Bluethgen, C., et al. (2024). A vision–language foundation model for the generation of realistic chest X-ray images. *Nature Biomedical Engineering, 9*(4), 494–506. https://doi.org/10.1038/s41551-024-01246-y

Chen, G., et al. (2025a). Elevating GPR imaging with DDPM: A comprehensive study on synthetic data generation. *IEEE Transactions on Geoscience and Remote Sensing, 63*, 1–18. https://doi.org/10.1109/TGRS.2025.3585865

Chen, Z., Guo, Y., Huang, Z., He, H., Lu, R., and Zhang, W., "High-quality proposal encoding and cascade denoising for imaginary supervised object detection", arXiv, Nov. 2025b. doi: 10.48550/arXiv.2511.08018.

Cui, L., et al. (2025). Advances of deep learning application in qualitative and quantitative detection of road subsurface distress using ground penetrating radar: A review. *Measurement, 247*, Article 116760. https://doi.org/10.1016/j.measurement.2025.116760

Dai, F., et al. (2025). Improving AI models for rare thyroid cancer subtype by text guided diffusion models. *Nature Communications, 16*(1), Article 4449. https://doi.org/10.1038/s41467-025-59478-8

Fan, Z., et al. (2025). RoadPrompt: A prompt engineering framework for road engineering large language models. *China Journal of Highway and Transport*. https://link.cnki.net/urlid/61.1313.u.20251212.1134.003.

Feng, D., Wang, X., & Zhang, B. (2018). Specific evaluation of tunnel lining multi-defects by all-refined GPR simulation method using hybrid algorithm of FETD and FDTD. *Construction and Building Materials, 185*, 220–229. https://doi.org/10.1016/j.conbuildmat.2018.07.039

Gulrajani, I., et al. (2017). Improved training of wasserstein GANs. *Advances in Neural Information Processing Systems, 30*, 5767–5777. https://proceedings.neurips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Heusel, M., et al. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems, 30*, 6626–6637. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems, 33*, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Hu, E., et al. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. https://iclr.cc/virtual/2022/poster/6319.

Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems, 35*, 26565–26577. https://doi.org/10.52202/068431-1926

Khosravi, B., Purkayastha, S., Erickson, B. J., Trivedi, H. M., & Gichoya, J. W. (2025). Exploring the potential of generative artificial intelligence in medical image synthesis: Opportunities, challenges, and future directions. *Lancet Digital Health, 7* (9), Article 100890. https://doi.org/10.1016/j.landig.2025.100890

Labs, B.F. et al. (2025), "FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space", arXiv:2506.15742, doi:10.48550/arXiv.2506.15742.

Lee, S., Lin, Z., & Fanti, G. (2024). Improving the training of rectified flows. *Advances in Neural Information Processing Systems*, 37, 63082–63109. https://doi.org/10.52202/079017-2014

Li, B., et al. (2023b). A grid-based classification and box-based detection fusion model for asphalt pavement crack. *Computer-Aided Civil and Infrastructure Engineering*. https://doi.org/10.1111/mice.12962

Li, F., Yang, F., Shi, W., Peng, S., & Li, Q. (Oct. 2025). GPR-VMD-Diffusion: A data generation algorithm using modes features of GPR signals for road underground targets. In *2025 IEEE Radar Conference (RadarConf25)* (pp. 1221–1226). https://doi.org/10.1109/RadarConf2559087.2025.11204943

Li, T., Xue, W., Yang, X., & Wu, D. (2023a). Road underground target detection in ground penetrating radar images based on YOLO models. In *China Automation Congress (CAC)* (pp. 2837–2842). https://doi.org/10.1109/CAC59555.2023.10450646. Nov.

Li, Y., et al. (2024). Adaptive curriculum learning for enhanced detection of subsurface distress using ground penetrating radar. *China Journal of Highway and Transport, 37* (12), 244–257. https://doi.org/10.19721/j.cnki.1001-7372.2024.12.007

Liu, S.-Y., et al. (2024). DoRA: Weight-decomposed low-rank adaptation. *Proceedings of Machine Learning Research, 235*, 32100–32121. https://proceedings.mlr.press/v235/liu24bn.html.

Liu, X., Gong, C., & Liu, Q. (2023). *Flow straight and fast: Learning to generate and transfer data with rectified flow. International Conference on Learning Representations*. https://openreview.net/forum?id=XVjTT1nw5z.

Liu, Z., Cui, B., & Gu, X. (2025). TIGPR: A multi-view ground penetrating radar detection data for damage assessment of transportation infrastructure. *Data in Brief, 60*, Article 111665. https://doi.org/10.1016/j.dib.2025.111665

Lv, H., Zhang, Y., Dai, J., Wu, H., Wang, J., & Wang, D. (2025). Intelligent recognition of GPR road hidden defect images based on feature fusion and attention mechanism. *IEEE Transactions on Geoscience and Remote Sensing, 63*, 1–17. https://doi.org/10.1109/TGRS.2025.3575293

Mojahid, A., Ouai, D. E., Amraoui, K. E., El-Hami, K., & Aitbenamer, H. (2025). Intelligent recognition of subsurface utilities and voids: A ground penetrating radar dataset for deep learning applications. *Data in Brief, 59*, Article 111338. https://doi.org/10.1016/j.dib.2025.111338

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. *Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, 70*, 2642–2651. https://proceedings.mlr.press/v70/odena17a.html.

Pham, M.-T., & Lefevre, S. (2018). Buried object detection from B-scan ground penetrating radar data using faster-RCNN. In *IEEE International Geoscience and Remote Sensing Symposium* (pp. 6804–6807). https://doi.org/10.1109/IGARSS.2018.8517683

Qin, H., Zhang, D., Tang, Y., & Wang, Y. (2021). Automatic recognition of tunnel lining elements from GPR images using deep convolutional networks with data augmentation. *Automation in Construction, 130*, Article 103830. https://doi.org/10.1016/j.autcon.2021.103830

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). Consistency models, Proceedings of the 40th International Conference on Machine Learning. *Proceedings of Machine Learning Research, 202*, 32211–32252. https://proceedings.mlr.press/v202/song23a.html.

Tang, X.-S., et al. (2026). Automatic detection of single-hyperbola pipeline signatures in urban roads using data-augmented ground penetrating radar profiles. *Results in Engineering, 29*, Article 108478. https://doi.org/10.1016/j.rineng.2025.108478

Wang, D., et al. (2023). Road structural defects detection and digitalization based on 3D ground penetrating radar technology: A state-of-the-art review. *China Journal of Highway and Transport, 36*(3), 1–19. https://doi.org/10.19721/j.cnki.1001-7372.2023.03.001

Wang, D., et al. (2025). Monitoring and detection technologies and AI-Powered development toward transparent roads. *Engineering.* , Article S2095809925008112. https://doi.org/10.1016/j.eng.2025.12.023

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *, 2. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers* (pp. 1398–1402). https://doi.org/10.1109/ACSSC.2003.1292216

Xiong, H., Li, J., Li, Z., & Zhang, Z. (2024). GPR-GAN: A ground-penetrating radar data generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 5200114. https://doi.org/10.1109/TGRS.2023.3337172

Yao, H., et al. (2025). Intelligent detection of subsurface road targets using a combined numerical simulation and deep learning method. *Computer-Aided Civil and Infrastructure Engineering, 40*(29), 5808–5820. https://doi.org/10.1111/mice.70121

Yue, G., Liu, C., Li, Y., Du, Y., & Guo, S. (2022). GPR data augmentation methods by incorporating domain knowledge. *Applied Sciences, 12*(21), Article 10896. https://doi.org/10.3390/app122110896

Zhang, H., Hu, Y., Wang, Q., Qian, Z., & Liu, P. (2025). End-to-end frequency enhancement framework for GPR images using domain-adaptive generative adversarial networks. *Computer-Aided Civil and Infrastructure Engineering, 40*(22), 3466–3482. https://doi.org/10.1111/mice.13525

Zhao, D., Guo, G., Ni, Z.-K., Pan, J., Yan, K., & Fang, G. (2023). WAEGAN: A GANs-based data augmentation method for GPR data. *IEEE Geoscience and Remote Sensing Letters, 20*, 1–5. https://doi.org/10.1109/LGRS.2023.3323981