Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Measurement-guided, training-free Fourier correction: a cost-efficient alternative to generative adaptation for cold-start construction-scene segmentation

Jonghun Gim

gimaiden19@jj.ac.kr

Jeonju University

Jeongik Min

Jeonju University

Research Article

Keywords: Cold-start semantic segmentation, Sim-to-real domain gap, Fourier domain adaptation, Synthetic training data, Construction-scene perception, Cost-efficient adaptation

Posted Date: July 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10340311/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Measurement-guided, training-free Fourier correction: a cost-efficient alternative to generative adaptation for cold-start construction-scene segmentation

Jonghun Gim[1*] and Jeongik Min[1]

[1*]Department of Artificial Intelligence, Jeonju University, Jeonju, 55069, Republic of Korea.

*Corresponding author(s). E-mail(s): gimaiden19@jj.ac.kr;
Contributing authors: minji@jj.ac.kr;

Abstract

Deploying pixel-level scene understanding at a new construction site is a cold-start problem: only a handful of labeled real frames are available, and the rare but safety- and task-critical classes—workers and dump piles—are precisely the hardest to learn. Labeled synthetic data can fill this gap, but the sim-to-real appearance mismatch limits its value, and the usual remedies, unsupervised domain adaptation and generative image translation, assume abundant unlabeled target data and compute that a cold-start site does not have. We introduce a measurement-guided correction protocol that (i) decomposes the low-frequency Fourier amplitude into global colour/exposure and texture-related components, (ii) quantifies each mismatch with sliced-Wasserstein distances, and (iii) verifies which component drives the downstream error before applying a single training-free pass of Fourier domain adaptation. On an excavator-worksite dataset with only 1% real labels, the protocol localizes the correctable mismatch to insufficient texture variation in simulated soil; correcting the low-frequency amplitude raises rare-class Dump IoU from 0.465 to 0.565 ($\boldsymbol{n = 6}$ seeds, $\boldsymbol{p < 0.001}$) by suppressing false-positive predictions on rough real regions. The training-free correction significantly outperforms generative translation with ControlNet, is statistically indistinguishable from far more expensive per-image diffusion restyling, and surpasses classical colour and augmentation corrections, at a fraction of their data and compute budget. The protocol is released as a reusable, measurement-first recipe for cold-start sim-to-real segmentation.

Keywords: Cold-start semantic segmentation, Sim-to-real domain gap, Fourier domain adaptation, Synthetic training data, Construction-scene perception, Cost-efficient adaptation

1 Introduction

Deploying pixel-level scene understanding at a new construction site is a cold-start problem. Dense annotation of excavator worksites is expensive, so a new site begins with only a handful of labeled real frames—on the order of 1% of a usable training set. The classes that matter most for autonomy and safety are the hardest to learn from so little data: workers (Human) are rare and safety-critical, and dump piles (Dump) are rare, task-relevant, and visually almost indistinguishable from ordinary soil. False Dump predictions on ordinary ground can misdirect autonomous path planning, so the practical objective is to reduce false positives for these rare classes without collecting more labels.

Labeled synthetic data from a simulator such as NVIDIA Isaac Sim can cheaply supply the missing rare-class examples, but rendered and real images differ, and this sim-to-real appearance gap limits the direct use of synthetic data [1, 2]. The dominant remedies are expensive. Unsupervised domain adaptation (UDA)—DAFormer, HRDA, DACS [3, 4, 11]—aligns feature or output distributions but assumes abundant unlabeled target images and full retraining. Generative image translation—ControlNet, DATUM [5, 12, 13]—produces target-like frames but needs a large data and compute budget, including per-site generator fine-tuning. A cold-start site has neither the unlabeled-target volume nor the compute these methods presuppose. This raises a question the adaptation literature rarely asks: what is the cheapest correction that is sufficient, and can it rival methods that cost orders of magnitude more?

We take a measurement-first position. Rather than fitting one adaptation model to the entire gap, we first measure which appearance component is worth correcting, then correct only that component with a single training-free pass. Concretely, we introduce a measurement-guided correction protocol (Fig. 1) that (i) decomposes the low-frequency Fourier amplitude into a global colour/exposure part and a texture-related part; (ii) quantifies each mismatch, and the residual structural mismatch, with sliced-Wasserstein distances over pixels and frozen deep features [6]; (iii) verifies the class and error pattern the mismatch produces with a confusion matrix and wrong-component/wrong-region controls; and (iv) applies the correction as a single training-free pass of Fourier domain adaptation (FDA) [2] that needs no generator, no target retraining, and no inference-time change. The protocol—not a new Fourier operator—is the contribution: it turns the choice of what to correct into an explicit, measured decision.

Instantiated on an excavator-worksite dataset with only 1% real labels, the protocol localizes the correctable mismatch to insufficient texture variation in simulated Soil, and shows that this—not global colour/exposure—is the effective lever. Correcting the low-frequency amplitude raises rare-class Dump IoU from 0.465 to 0.565 ($+0.100$; $n = 6$ seeds; $p < 0.001$), driven by fewer false-positive Dump predictions on

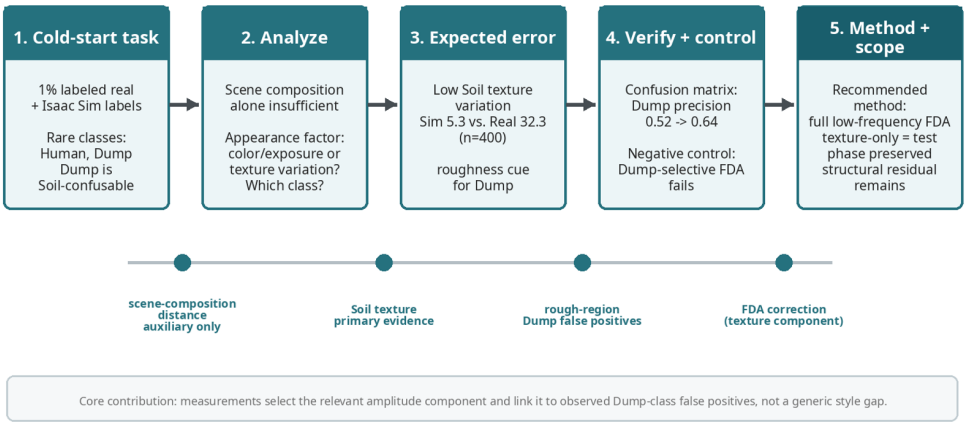


Fig. 1 Overview of the measurement-guided correction protocol. The five stages run left to right: cold-start task definition (1% labeled real frames plus Isaac Sim labels), gap analysis (scene composition vs. appearance), expected-error derivation (low simulated Soil texture variation induces a roughness cue for Dump), verification with a confusion matrix and a negative control, and the recommended training-free correction (full low-frequency FDA; the texture-only variant serves as the mechanistic probe). The core contribution is that measurements select the amplitude component to correct and link it to the observed Dump-class false positives

rough real regions. Our central result is one of cost-efficiency: this training-free correction matches or exceeds substantially more expensive image-correction strategies at a fraction of their data and compute budget. It significantly outperforms generative translation (ControlNet, +0.036, $p = 0.022$; DATUM is statistically indistinguishable from raw synthetic data), and classical colour/augmentation baselines (histogram matching, Reinhard transfer, photometric jitter) also underperform, confirming the gain is texture-specific rather than generic realism.

Contributions. (C1) We propose a measurement-guided, training-free correction protocol that selects the appearance component to correct from explicit distributional measurements, rather than adapting the whole gap blindly. (C2) We show that, in the 1% cold-start regime, this cheap correction is Pareto-competitive with or superior to generative image translation at a small fraction of the cost. (C3) We identify the mechanism—a texture-variance deficit in simulated Soil that induces false-positive Dump predictions—and verify it with confusion-matrix and negative controls. (C4) We bound the method's reach: amplitude correction closes 53% of the pixel-level Soil gap but only ∼15–30% of the deep-feature gap, so the residual is largely structural. We validate on one construction case and release the protocol as a reusable, measurement-first recipe; broader validation across sites and class pairs is future work.

2 Related work

2.1 Fourier and amplitude-spectrum manipulation

FDA [2] transfers low-frequency amplitude information from target to source images while preserving phase. Style-transfer and amplitude-spectrum methods [7, 8, 9, 10] also modify appearance statistics, but they do not specify which spectral component causes a class-specific error. Here, FDA is used as both an analysis and correction tool: we compare global colour/exposure with the texture-related component of the low-frequency amplitude window and link the effective component to the measured Soil mismatch and verified false-positive Dump predictions.

2.2 UDA and generative adaptation, and why they are out of scope here

UDA methods such as DAFormer, HRDA, and DACS [3, 4, 11] align feature representations or use cross-domain mixed sampling, but they usually require abundant unlabeled target images and full retraining. Generative adaptation methods such as DATUM and ControlNet [5, 12, 13] create target-like synthetic images but require larger data and compute budgets. Our setting uses only amplitude spectra from the 1%-labeled real training frames in a single offline pass. We therefore treat UDA as complementary rather than a competing baseline: in the extreme cold-start regime studied here, the unlabeled-target volume and per-site retraining that UDA presupposes are not available, and a fair, budget-matched UDA comparison is itself a separate study (Section 7.3). We instead benchmark against the image-correction strategies that *are* applicable at cold start—generative translation (ControlNet, DATUM) and classical colour/augmentation corrections—which share our "modify the synthetic training images" interface.

2.3 Synthetic data and domain randomization

Synthetic datasets and domain randomization [1, 14, 15, 16] reduce annotation cost by randomizing rendered appearance. In contrast, we first measure the discrepancy and then modify only the component linked to the downstream segmentation error.

2.4 Low-label and few-shot segmentation

Few-shot and semi-supervised segmentation methods [17, 18, 19] use support sets or unlabeled real images. Our additional supervision is labeled synthetic data; the central problem is therefore the synthetic-to-real appearance mismatch.

2.5 Construction-scene perception

Construction-site vision has largely focused on object detection, personal protective equipment detection, and safety monitoring [20]. Dense segmentation for excavator worksites under severe label scarcity, especially Soil–Dump confusion, remains less studied.

3 Problem formulation and experimental setup

3.1 The sim-to-real gap as a band-wise discrepancy

We define a measurable proxy for the sim-to-real domain gap. Let synthetic and real image distributions be given, and let $D(\cdot)$ be a sliced-Wasserstein distance [6] computed on pixel intensities and frozen deep features (Section 7.1). We decompose the Fourier amplitude A into global colour/exposure and texture-related components within the low-frequency window:

$$A = A_{\text{colour/exposure}} + A_{\text{texture}} + A_{\text{residual}}. \tag{1}$$

The first two terms denote measurable appearance contributions, while the residual term corresponds to phase, structure, and geometry, which amplitude editing cannot change. The analysis estimates which term in Eq. (1) is large and correctable. FDA [2] then replaces the selected source-amplitude band with amplitude statistics from a small real corpus while preserving phase. The texture-only variant replaces only the nonzero part of the low-frequency window, leaving global colour/exposure and labels unchanged. We use D as a task-relevant proxy for the discrepancy term in the domain-adaptation bound [24]:

$$\varepsilon_{\text{target}} \leq \varepsilon_{\text{source}} + D(\text{synthetic}, \text{real}) + \lambda. \tag{2}$$

If this proxy captures a task-relevant mismatch, synthetic frames corrected by the selected amplitude transfer and mixed with 1% labeled real frames should improve rare-class segmentation (Section 5). This is an empirical hypothesis: frequency components can interact, the full low-frequency swap is not additive (Section 5.2), and the phase/structure residual explains why the correction closes much of the pixel-level gap but only part of the deep-feature gap (Section 7.1).

3.2 Data, task, and cold-start protocol

Data and task. Real images are from the AIHub construction-machinery corpus, and synthetic images are rendered in Isaac Sim using an asset-hybrid pipeline. We segment Soil, Self, Human, Dump, Obstacle, and Background, plus an ignore label. Real frames are dominated by Soil (0.52), Obstacle (0.29), and Self (0.17); Human and Dump occupy very small areas and are difficult to segment.

Cold-start protocol. We use 1% of real frames as labels, optionally augmented with synthetic frames at an approximately 50% ratio (real-prior class weighting; synthetic loss $\times 0.5$).

Data budget and leakage control. The recommended global correction uses amplitude statistics from 300 real frames sampled from the 1%-labeled training set. It uses only full-image amplitude spectra, not labels, masks, or additional real frames. No validation or test frame is used. By contrast, the generative baselines (Section 5.4) require extra real data and training time, including DATUM fine-tuning on 50 real frames. Our method therefore uses a smaller real-data and compute budget.

3.3 Models, metric, and evaluation

Models. The main model is a from-scratch Residual U-Net [21]; we also use ImageNet-pretrained SegFormer-B2 [22] for a backbone comparison. The loss is weighted cross-entropy plus Dice loss [23]. All comparisons use the same splits and protocol.

Training details. All models are trained at 1024×576 resolution with AdamW (learning rate 10^{-4}, weight decay 10^{-4}), a cosine-annealed schedule decaying to 10^{-6}, batch size 4 with two-step gradient accumulation, and mixed precision, for up to 60 epochs with early stopping (patience 15). Mixed batches are drawn by a weighted sampler at an approximately 70:30 real-to-synthetic ratio, and synthetic pixels contribute to the loss with weight 0.5, so that the sparse real labels remain the primary supervision signal. Random flip augmentation is applied during training, and the final evaluation of each run uses flip test-time augmentation. All matched comparisons use the same six fixed random seeds, identical data manifests, and an identical checkpoint-selection rule, so paired statistics compare runs that differ only in the synthetic-image correction applied.

Evaluation. We evaluate on val_cross_sequence, a sequence-disjoint validation split. A same-sequence split is optimistic because adjacent video frames are near-duplicates; for example, the same model obtains Dump IoU 0.96 on val_official but 0.56 on val_cross_sequence. We report IoU, precision, recall, and mIoU, emphasizing Dump (rare and similar to Soil) and Human (safety critical). Checkpoints are selected by val_cross_sequence mIoU; absolute values should be interpreted as sequence-disjoint validation estimates, with the same selection rule applied to all methods.

4 Gap analysis

Equation (1) asks which correctable appearance component—global colour/exposure, texture-related amplitude information, or neither—explains the task-relevant discrepancy. We address this by comparing scene composition with appearance, measuring within-frame surface variation, and verifying the corresponding error pattern with a confusion matrix and negative control.

4.1 Scene-composition alignment alone is insufficient

We first test whether changing scene composition—class proportions and object placement—is sufficient. Four synthetic strategies are added to the 1% real baseline. Under a multi-metric criterion (one seed from an earlier phase-1 series), none yields a Pareto improvement. Distribution matching gives the highest Dump IoU in that screen (0.596) but reduces Human IoU from 0.429 to 0.155. A mask-only sliced-Wasserstein distance over class composition and coarse spatial occupancy spans a meaningful range of alignment with real data (1787–2526), yet the distance does not track Dump true-positive segmentation: the closest composition gives one of the lowest Dump IoUs (0.446), whereas the farthest gives the highest (0.596). Scene-composition alignment alone is therefore insufficient; appearance evidence is provided in Sections 4.2–5.

4.2 Simulated Soil has insufficient surface variation

We localize the appearance gap to surface variation while controlling a measurement confound. Per-class intensity standard deviation mixes within-frame texture with cross-frame brightness/exposure variation, so the two sources are analyzed separately (Fig. 2).

The central finding concerns Soil ($n = 400$). Simulated Soil has much lower within-frame luminance variation than real Soil (5.3 vs. 32.3), and the correction moves this value toward the real value (26.7). A naive pooled statistic would make simulated Dump appear over-textured, but this is caused by cross-frame brightness variation (simulated 56.2 vs. real 5.3), not surface roughness. The likely rendering cause is a near-uniform ground albedo and limited micro-geometry, which remove fine intensity variation from simulated Soil.

Mechanism and expected confusion pattern. In the simulator, Soil has lower within-frame variation than Dump piles (5.3 vs. 17.9). A model may therefore learn roughness as a spurious cue for Dump (Fig. 3). In real images, rough regions include soil, obstacles, and machine parts, so the cue can produce false-positive Dump predictions and lower precision. We evaluate this expected pattern with a column-normalized confusion matrix.

4.3 Error verification and negative control

Verification. Using a column-normalized confusion matrix (precision view), in the baseline only 52% of predicted Dump pixels are true Dump; most false positives come from rough Obstacle (20%) and Self (20%) regions, while Soil→Dump is smaller (8%). After texture-related amplitude correction, Dump precision increases to 64% ($n = 6$ matched seeds), mainly because Obstacle→Dump and Self→Dump errors decrease ($0.20 \rightarrow 0.13$ and $0.20 \rightarrow 0.14$; Fig. 4). Human precision also increases from 0.72 to 0.78. A pixel-intensity Wasserstein check gives the same direction: the Soil discrepancy decreases from 0.133 to 0.063 (-53%).

Negative control. To confirm that the gain is not produced by arbitrary FDA, we apply the same correction to an intentionally inappropriate region using Dump-selective FDA. This variant does not improve the result (Dump IoU 0.547, a Pareto failure; one seed), whereas the global correction does.

4.4 The measurement-based protocol

We state the procedure as a protocol (Algorithm 1) to emphasize measurement-based target selection rather than a new Fourier operator.

5 Method: measurement-guided, training-free Fourier amplitude correction

The analysis in Section 4 identifies the Soil texture-related amplitude mismatch as the dominant correctable term in Eq. (1). Section 5.1 defines the correction; Section 5.2 shows the texture-related component is responsible for the gain; Section 5.3 reports the

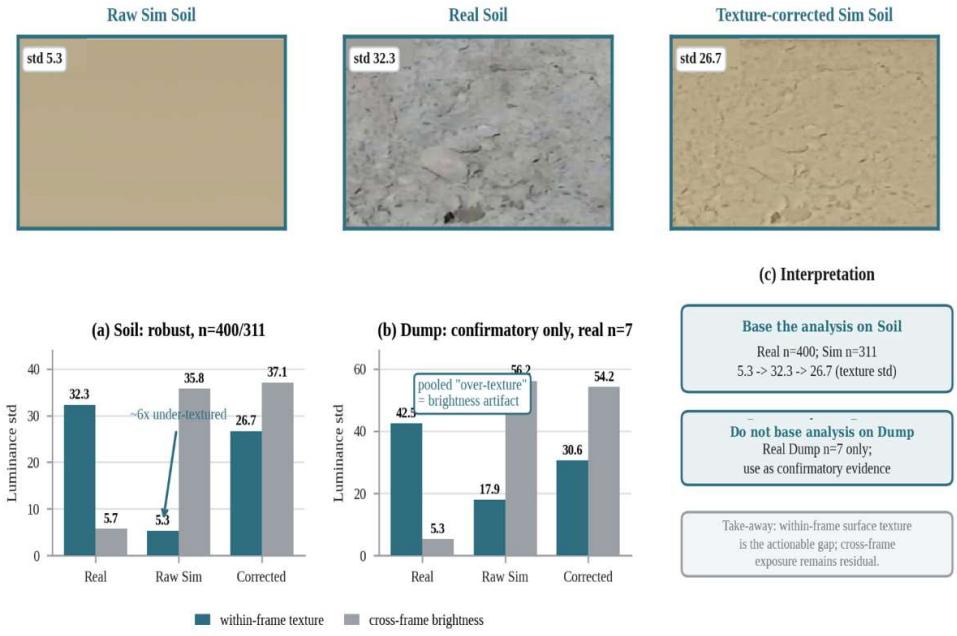


Fig. 2 Within-frame texture vs. cross-frame brightness decomposition of per-class luminance variation. **a** Soil (robust, real $n = 400$, simulated $n = 311$): simulated Soil is $\sim$6$\times$ under-textured relative to real Soil (within-frame std 5.3 vs. 32.3), and the texture-corrected synthetic data moves it toward the real value (26.7). **b** Dump (confirmatory only, real $n = 7$): the apparent over-texture of simulated Dump in a pooled statistic is a cross-frame brightness artifact (56.2 vs. 5.3). **c** Interpretation: the analysis is based on Soil; Dump statistics are used only as confirmatory evidence. Example patches (top) show raw simulated, real, and texture-corrected simulated Soil

cold-start improvement; and Section 5.4 compares heavier learned image-correction baselines.

5.1 Method

We start from global FDA [2], which replaces source low-frequency amplitude with target-domain amplitude while preserving phase. The correction is a single offline pass over synthetic frames; it requires no generator training or inference-time model change.

Formally, let $\mathcal{F}$ denote the 2-D discrete Fourier transform, and let $A_s = |\mathcal{F}(x_s)|$ and $\phi_s = \angle\mathcal{F}(x_s)$ be the amplitude and phase of a synthetic frame x_s. With M_β a centred low-frequency window covering a fraction β of the spectrum, the corrected frame is

$$\hat{x}_s = \mathcal{F}^{-1}\left[(1 - M_\beta) \odot A_s e^{i\phi_s} + M_\beta \odot \hat{A}_t e^{i\phi_s}\right], \tag{3}$$

where $\hat{A}_t$ is the amplitude statistic estimated from the 300-frame real corpus. Phase—and therefore structure, geometry, and label alignment—is untouched, so the segmentation labels of the synthetic frames remain valid. Within the low-frequency window, the zero-frequency (DC) coefficient carries global colour and exposure,

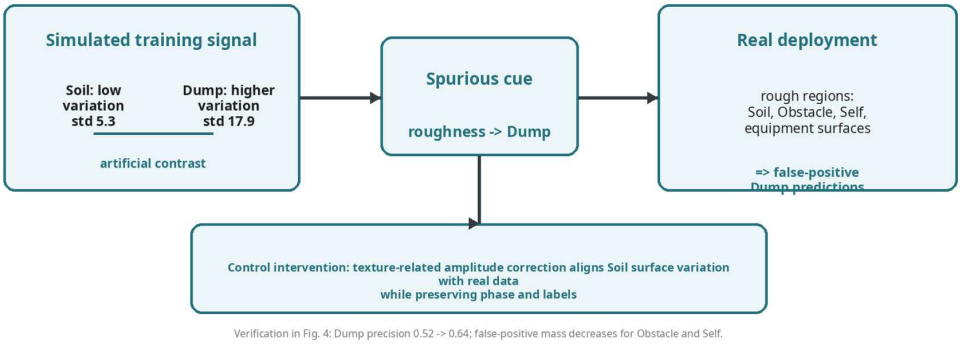


Fig. 3 Hypothesized error mechanism. The simulator presents an artificial contrast between low-variation Soil (std 5.3) and higher-variation Dump (std 17.9), so a model can learn roughness as a spurious cue for Dump. At deployment, rough real regions (soil, obstacles, machine surfaces) then attract false-positive Dump predictions. The control intervention aligns simulated Soil surface variation with real data through texture-related amplitude correction while preserving phase and labels; verification appears in Fig. 4

Algorithm 1 Measurement-guided Fourier amplitude correction

Require: Labeled synthetic data, a small real corpus, a from-scratch segmenter
Ensure: Selected correction target and corrected synthetic set

1: Test scene-composition alignment.
2: Split the low-frequency amplitude window into global colour/exposure and texture-related components; estimate discrepancies and run colour-only, texture-only, and full-amplitude ablations.
3: Identify the affected class and dominant mismatch.
4: Verify the confusion pattern and apply a wrong-component or wrong-region control.
5: Replace the selected source-amplitude component with real-corpus statistics, preserve phase, and train with corrected synthetic data plus 1% labeled real data.

while the remaining low-frequency (AC) coefficients carry coarse surface texture; the colour/exposure-only, texture-only, and full variants of Section 5.2 correspond to swapping the DC part, the AC part, or both. The correction has no learned parameters and costs one forward and one inverse FFT per synthetic frame. We set $\beta = 0.05$ a priori in the small-value regime recommended for FDA and keep it fixed, so it is not tuned on val_cross_sequence. The contribution is measurement-based component selection: Section 4 localizes the correctable mismatch to Soil surface variation, so the motivated intervention is to replace the texture-related component. The full low-frequency amplitude swap is the recommended deployment form because it is most robust to checkpoint selection (Section 5.3); the texture-only swap is the mechanistic test. Soil-region-only FDA gives no additional benefit over the global form (Dump IoU $\Delta+0.01$, paired $p = 0.80$, $n = 3$), so the method uses the simpler global correction.

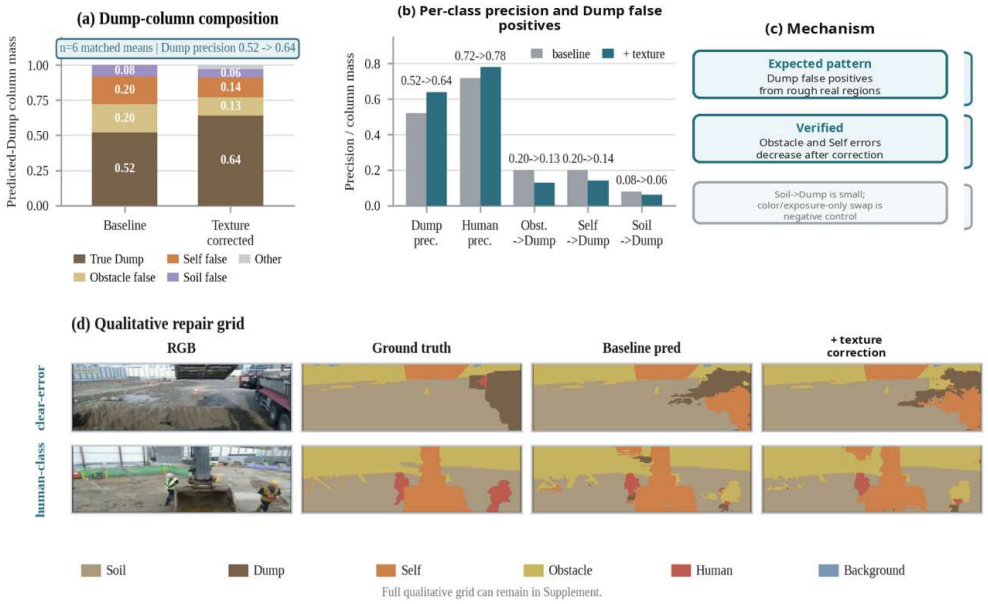


Fig. 4 Verification of the predicted false-positive pattern ($n = 6$ matched seeds). **a** Composition of the predicted-Dump column: true-Dump share rises from 0.52 to 0.64 after texture correction. **b** Per-class precision and Dump false-positive shares: Obstacle→Dump (0.20 → 0.13) and Self→Dump (0.20 → 0.14) decrease, Human precision rises (0.72 → 0.78), and the small Soil→Dump share also decreases (0.08 → 0.06). **c** Mechanism summary: the expected pattern (false positives from rough real regions) is verified. **d** Qualitative repair grid: RGB, ground truth, baseline prediction, and prediction with texture correction for a clear-error case and a human-class case

5.2 Colour/exposure vs. texture-related amplitude: an interventional ablation

We split the transferred low-frequency amplitude into global colour/exposure and texture-related components and train three variants—colour/exposure only, texture only, and full low-frequency transfer—using the same six matched seeds (Table 1). The texture-only swap matches the full swap (Dump IoU 0.568 vs. 0.565, paired $\Delta + 0.003$, $p \approx 0.84$), gives the highest rare-class precision (0.643), and has the lowest variance ($\sigma = 0.021$). The colour/exposure-only swap remains near baseline (Dump IoU 0.472 vs. 0.465; precision 0.517 vs. 0.520) and is the most unstable variant ($\sigma = 0.080$, range 0.368–0.594), consistent with clipping from isolated colour/exposure transfer that reduces the texture signal of Dump regions. A column-normalized confusion analysis supports the same conclusion: the Soil→Dump false-positive share is 0.077 at baseline, 0.060 with texture-only transfer, 0.068 with full transfer, and 0.158 with colour/exposure-only transfer. In this case, texture-related amplitude information explains the improvement; the broader contribution is the procedure for identifying the relevant factor, which may select a different component in another domain.

Table 1 Interventional ablation of the transferred amplitude components ($n = 6$ matched seeds). The texture-related (AC) component, selected by the measurements of Section 4, explains the improvement; transferring colour/exposure (DC) alone is ineffective and unstable

Variant	Dump IoU	Dump precision	Soil→Dump FP share
Baseline (1% real + raw synthetic)	0.465	0.520	0.077
Colour/exposure-only swap	0.472	0.517	0.158
Texture-only swap	0.568	0.643	0.060
Full low-frequency swap (deployed)	0.565	0.630	0.068

Table 2 Cold-start segmentation results with the full low-frequency amplitude correction ($n = 6$ matched seeds; paired t-tests). The gain is concentrated in the rare Dump class and is driven by precision; majority-class performance is unchanged

Metric	Baseline	+Full correction	p
Dump IoU	0.465	0.565	0.0007
Dump precision	0.52	0.63	–
Dump recall	0.82	0.85	–
Human IoU	0.507	0.541	0.115
Soil IoU	0.956	0.956	–
mIoU	0.755	0.778	< 0.0001

5.3 Cold-start improvement

In the 1% cold-start regime, global low-frequency amplitude correction improves Dump IoU from 0.465 to 0.565 (+0.100; paired t-test $p = 0.0007$, $d_z = 3.0$; 6/6 seeds improve; Fig. 5; Table 2). mIoU increases from 0.755 to 0.778 (+0.023; $p < 0.0001$), Human IoU increases from 0.507 to 0.541 (+0.034; $p = 0.115$), and Soil IoU is unchanged ($0.956 \rightarrow 0.956$). The gain is concentrated in the rare Dump class. Because the +0.10 value uses the best val_cross_sequence checkpoint, a selection-free check reduces the gain to about +0.07 for the full swap and +0.04 for the texture-only swap, indicating that the full swap is more robust. Only one of the two held-out sequences contains Dump, but both sequences show fewer pixels incorrectly predicted as Dump: −30% on the Dump-containing sequence and −43% on the Dump-free sequence. The method raises Dump precision from 0.52 to 0.63 while maintaining recall ($0.82 \rightarrow 0.85$). Seed-level p-values measure stability across random initializations, not generalization across sites; a sequence-level paired bootstrap is inconclusive with only two sequences (Δ Dump IoU +0.078, 95% CI $[-0.00, 0.11]$). For safety-critical deployment, method selection should include an explicit Human-recall constraint, since the full method increases Human precision (+0.088) but slightly lowers recall (−0.013).

5.4 Comparison with generative and classical image correction (cost-efficiency)

We compare against ControlNet [13] (Stable-Diffusion-1.5 with Canny conditioning) and DATUM [5, 12] (50-image DreamBooth fine-tuning of SD-1.5, img2img strength

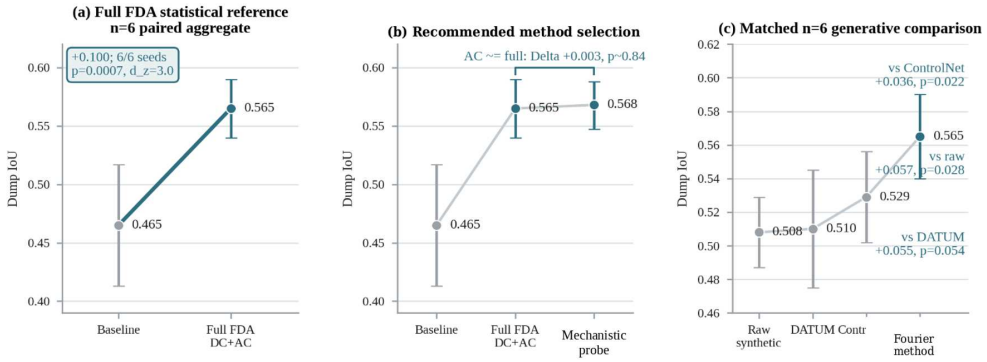


Fig. 5 Cold-start results ($n = 6$ matched seeds; error bars show seed variability). **a** Full low-frequency FDA improves Dump IoU from 0.465 to 0.565 (+0.100; 6/6 seeds; $p = 0.0007$, $d_z = 3.0$). **b** Recommended method selection: the texture-only mechanistic probe (0.568) is statistically equivalent to the full swap ($\Delta + 0.003$, $p = 0.84$); the full swap is recommended for deployment because it is more robust to checkpoint selection. **c** Matched generative comparison: the Fourier method (0.565) outperforms raw synthetic data (+0.057, $p = 0.028$) and ControlNet (+0.036, $p = 0.022$), while DATUM (0.510) is statistically indistinguishable from raw synthetic data (0.508)

Table 3 Matched comparison of image-correction strategies at cold start ($n = 6$ seeds). Δ and p are paired differences of the proposed full amplitude swap over each alternative. The proposed correction uses only amplitude spectra of 300 frames already contained in the 1%-labeled training set, requires no generator, and runs as a single offline pass

Method	Dump IoU	Δ ours (p)	Generator training	Correction cost
Raw synthetic	0.508	+0.057 (0.028)	none	none
ControlNet (Canny, SD-1.5)	0.529	+0.036 (0.022)	none (pretrained)	per-image diffusion
DATUM (DreamBooth, SD-1.5)	0.510	+0.055 (0.054)	per-site fine-tuning[1]	per-image img2img
Ours (full amplitude swap)	0.565	–	none	one offline FFT pass

[1]DreamBooth fine-tuning of Stable Diffusion 1.5 on 50 additional real frames.

0.5) under the same 1%-real cold-start setting and $n = 6$ matched seeds. Each method modifies the synthetic training images and trains the same from-scratch Residual U-Net; this section therefore compares image-correction strategies, not segmentation architectures. "Ours" denotes the full low-frequency amplitude swap (Dump IoU 0.565); the texture-only swap is statistically equivalent (0.568).

With $n = 6$ matched seeds, the proposed method gives the highest Dump IoU: +0.057 over raw synthetic data ($p = 0.028$) and +0.036 over ControlNet ($p = 0.022$); DATUM (0.510) is statistically indistinguishable from raw synthetic data (0.508), suggesting that per-image diffusion restyling may distort structure rather than add the missing surface variation. Classical controls confirm that the gain is not simply due to generic colour realism or added high-frequency energy: histogram matching, Reinhard colour transfer, and photometric jitter all underperform the Fourier-domain correction.

Crucially, our method attains this while using no generator, no fine-tuning, and a single offline pass—orders of magnitude less data and compute than the generative baselines. The result is best read as cost-efficiency: at cold start, a measurement-selected training-free correction is Pareto-competitive with or superior to substantially more expensive image-correction pipelines.

6 Operating boundaries and practical ceiling

We identify conditions under which the method provides limited benefit. Class-selective correction is unnecessary (paired $p = 0.80$, $n = 3$; preliminary negative result). With a pretrained SegFormer-B2 backbone, the gain shrinks to +0.022 Dump IoU, consistent with pretraining absorbing much of the appearance mismatch. The $\leq \approx 2\%$ annotation boundary is based on one seed and is directional. The correction gives a large relative gain for the rare Dump class (+21.5%) under a small real-data budget, while changes in mIoU and Human performance are modest because the method targets one measured error pattern rather than redesigning the segmentation system.

7 Discussion

7.1 How much of the gap is explained by appearance?

Equation (1) includes a phase/structure residual that amplitude editing cannot directly change. We measure the appearance-related share of the representational gap with learning-free sliced-Wasserstein distance over frozen deep features from ResNet50 and VGG16 (Fig. 6). At the pixel level, the correction closes 53% of the Soil gap ($n = 400$; W_1 0.133 $\rightarrow$ 0.063) and 66% of the Dump gap (W_1 0.154 $\rightarrow$ 0.052). At intermediate deep representations, both architectures indicate only $\sim$15–30% closure. Appearance is therefore a partial contributor to the deep-feature gap: FDA preserves phase, which contains much of the structure, geometry, and layout, and edits amplitude, which mainly affects appearance. Appearance correction is a high-return first step, but closing the majority residual will require higher-fidelity simulation or structure-level adaptation.

7.2 Limitations

The primary claims are the Soil texture-variation analysis ($n = 400$) and the matched-seed Dump-IoU improvement. Real Dump texture statistics are underpowered ($n = 7$), some boundary studies are single-seed, and the results cover one dataset, simulator, and class pair. The colour/exposure result concerns amplitude-level transfer with clipping and is not a general claim about colour information.

7.3 Scope and future work: relation to UDA

We deliberately position UDA as complementary rather than as a head-to-head baseline. Modern UDA (DAFormer, HRDA, DACS) is designed for a different budget—abundant unlabeled target frames and full retraining—whereas our contribution is a

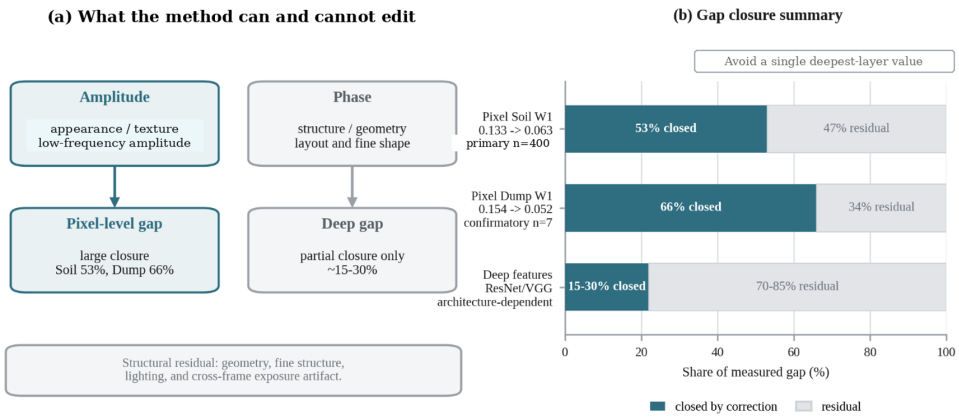


Fig. 6 Scope of amplitude correction. **a** What the method can and cannot edit: amplitude carries appearance (low-frequency amplitude), whereas phase carries structure, geometry, and layout; the structural residual includes geometry, fine structure, lighting, and cross-frame exposure artifacts. **b** Gap-closure summary: pixel-level sliced-Wasserstein distances close by 53% (Soil, $n = 400$) and 66% (Dump, confirmatory $n = 7$), while frozen deep-feature distances (ResNet50/VGG16) close by only ∼15–30%

single offline correction usable when only 1% of frames are labeled and neither a large unlabeled-target pool nor per-site retraining is practical. A budget-matched UDA comparison (e.g., DACS on the same backbone with only the unlabeled frames available at cold start) is a natural and important extension, and we expect the two directions to be combinable: a measurement-selected training-free correction as a cheap first step, followed by UDA when a larger unlabeled-target pool and compute become available.

7.4 Practical guidance

Synthetic simulator data are most useful when real labels are limited to $\leq \approx 2\%$, the backbone is trained from scratch, and the ablation identifies texture-related amplitude information as the dominant correctable mismatch. Returns diminish when the remaining gap is structural or a strong pretrained backbone already reduces the appearance mismatch.

In deployment terms, the recommended recipe is: collect the already-available 1% labeled frames; estimate the corpus amplitude statistic once; apply Eq. (3) to every synthetic frame in a single offline pass; and train the site model as usual. The correction adds no inference-time cost and no new run-time failure modes, because the deployed network is architecturally unchanged; the entire overhead is a one-off preprocessing step of one forward and one inverse FFT per synthetic frame. When a site later accumulates a larger pool of unlabeled frames, the corrected synthetic set remains a valid starting point for subsequent unsupervised domain adaptation (Section 7.3).

8 Conclusion

We presented a measurement-guided, training-free correction protocol for cold-start construction-scene segmentation. The protocol identifies the correctable appearance component before applying the correction; in this case, texture-related low-frequency amplitude information in Soil regions reduces false-positive Dump predictions on rough real regions and improves the rare Dump class. Its central value is cost-efficiency: a single offline pass matches or outperforms substantially more expensive generative image correction, while classical corrections underperform. Because the evidence comes from a single site, future work should apply the same protocol to other class pairs, sites, and simulators, and combine it with budget-matched unsupervised domain adaptation.

Declarations

Funding. This research received no specific grant from public, commercial, or not-for-profit funding agencies.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. This study uses a publicly distributed, de-identified AIHub dataset; no additional human-subjects data were collected, and no separate institutional review was required.

Consent for publication. Not applicable.

Data availability. The real images are from the AIHub Unmanned Operation Data in Construction Machinery dataset (dataset no. 081-102), distributed by AIHub (https://www.aihub.or.kr) under its terms of use; the authors cannot redistribute the raw data. Synthetic renders and generation configurations are available from the authors upon reasonable request.

Materials availability. Not applicable.

Code availability. The analysis scripts and the training-free Fourier-domain correction procedure are available from the corresponding author upon reasonable request and will be released with configuration files and fixed seeds upon acceptance.

Author contribution. Jonghun Gim: conceptualization, methodology, experiments and analysis, writing—original draft. Jeongik Min: supervision, validation, writing—review and editing. All authors read and approved the final manuscript.

Use of generative AI. A large-language-model-based assistant was used for language editing and formatting. The authors reviewed and verified all content and take responsibility for the scientific content, analyses, and conclusions.

References

[1] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real

world. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp. 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133
[2] Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4084–4094 (2020). https://doi.org/10.1109/CVPR42600.2020.00414
[3] Hoyer, L., Dai, D., Van Gool, L.: DAFormer: improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 9914–9925 (2022). https://doi.org/10.1109/CVPR52688.2022.00969
[4] Hoyer, L., Dai, D., Van Gool, L.: HRDA: context-aware high-resolution domain-adaptive semantic segmentation. In: Proc. European Conf. Computer Vision (ECCV), pp. 372–391 (2022). https://doi.org/10.1007/978-3-031-20056-4_22
[5] Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., Lathuilière, S.: One-shot unsupervised domain adaptation with personalized diffusion models. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 698–708 (2023). https://doi.org/10.1109/CVPRW59228.2023.00077
[6] Bonneel, N., Rabin, J., Peyré, G., Pfister, H.: Sliced and Radon Wasserstein barycenters of measures. J. Math. Imaging Vis. **51**(1), 22–45 (2015). https://doi.org/10.1007/s10851-014-0506-3
[7] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
[8] Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A Fourier-based framework for domain generalization. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 14378–14387 (2021). https://doi.org/10.1109/CVPR46437.2021.01415
[9] Yang, Y., Wang, S., Zhu, L., Yu, L.: HCDG: a hierarchical consistency framework for domain generalization on medical image segmentation. arXiv preprint arXiv:2109.05742 (2021)
[10] Chattopadhyay, P., Sarangmath, K., Vijaykumar, V., Hoffman, J.: PASTA: proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 19231–19243 (2023). https://doi.org/10.1109/ICCV51070.2023.01767
[11] Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: domain adaptation via cross-domain mixed sampling. In: Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), pp. 1379–1389 (2021). https://doi.org/10.1109/WACV48630.2021.00142
[12] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine-tuning text-to-image diffusion models for subject-driven generation. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510 (2023). https://doi.org/10.1109/CVPR52729.2023.02155
[13] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 3836–3847 (2023). https://doi.org/10.1109/ICCV51070.2023.00355

[14] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016). https://doi.org/10.1109/CVPR.2016.352
[15] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: ground truth from computer games. In: Proc. European Conf. Computer Vision (ECCV), pp. 102–118 (2016). https://doi.org/10.1007/978-3-319-46475-6_7
[16] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350
[17] Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: Proc. British Machine Vision Conference (BMVC) (2017). https://doi.org/10.5244/C.31.167
[18] Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANet: few-shot image semantic segmentation with prototype alignment. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 9196–9205 (2019). https://doi.org/10.1109/ICCV.2019.00929
[19] Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 12671–12681 (2020). https://doi.org/10.1109/CVPR42600.2020.01269
[20] Nath, N.D., Behzadan, A.H., Paal, S.G.: Deep learning for site safety: real-time detection of personal protective equipment. Autom. Constr. **112**, 103085 (2020). https://doi.org/10.1016/j.autcon.2020.103085
[21] Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
[22] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 12077–12090 (2021)
[23] Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proc. Int. Conf. 3D Vision (3DV), pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
[24] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010). https://doi.org/10.1007/s10994-009-5152-4