Active Sim-to-Real Gap Reduction for Industrial Inspection via Digital Twin and Embedding Analysis

Huimin Zhuge[1], Xian Yeow Lee[1], Gregory Sin[1], Raheem Ahmed[1], Lasitha Vidyaratne[1], Aman Kumar[1], and Ahmed Farahat[1]

[1] *Hitachi America Ltd, Santa Clara, California, USA*
joy.zhuge@hal.hitachi.com
xian.lee@hal.hitachi.com
gregory.shobe@hal.hitachi.com
raheem.ahmed@hal.hitachi.com
lasitha.vidyaratne@hal.hitachi.com
aman.kumar@hal.hitachi.com
ahmed.farahat@hal.hitachi.com

ABSTRACT

Simulation-based training is increasingly used in automated industrial inspection, where collecting and annotating real-world inspection data is costly and often impractical. While synthetic data generated from digital twins enables scalable training, models trained solely in simulation suffer from a significant sim-to-real gap under real inspection conditions such as varying lighting, surface properties, and sensor noise. In this work, we propose a data-efficient sim-to-real adaptation framework that combines representative sample selection via k-determinantal point processes (k-DPP) with embedding-level alignment using Kullback–Leibler (KL) divergence. The key idea is to actively identify a small set of representative synthetic samples, acquire the corresponding real images, and align their latent feature representations while retaining the coverage provided by the larger synthetic dataset. We first train an RF-DETR(Detection Transformer) detector on 550 synthetic inspection images, achieving near-perfect performance in simulation but only 0.2516 mean Average Precision (mAP) on real-world images. Using only 50 paired real images (approximately 10% of the synthetic training set) together with 500 unpaired synthetic images, the proposed method increases real-world mAP from 0.2516 to 0.8853. The k-DPP sampling strategy maximizes the diversity of selected samples, reducing the risk of bias introduced by limited real-world data, while KL-based embedding alignment further reduces domain discrepancy between synthetic and real images. The proposed framework provides a lightweight and practical approach for reducing sim-to-real gaps in a representative industrial inspection setting where real data collection is limited.

1. INTRODUCTION

Object detection systems have achieved significant progress with the development of deep learning models and large-scale annotated datasets. Modern architectures such as transformer-based detectors have demonstrated strong performance across many visual recognition tasks (Carion et al., 2020). However, collecting and labeling real-world training data remains expensive and time-consuming, particularly in industrial environments where data acquisition may require specialized equipment or controlled conditions. As a result, synthetic data generated from simulation environments or digital twins has become an attractive alternative for training vision models.

Synthetic datasets allow scalable data generation, precise annotations, and flexible control over environmental conditions such as lighting, camera viewpoints, and object placement. Despite these advantages, models trained exclusively on synthetic data often fail to generalize well to real-world scenarios. This challenge, commonly referred to as the sim-to-real gap, arises from discrepancies between synthetic and real images, including differences in texture, lighting, noise, and material reflectance. Even when the geometric structure of objects is accurately modeled, subtle appearance variations can significantly affect model performance.

Several approaches have been proposed to address the sim-to-real gap. One widely used strategy is domain randomization, which introduces extensive variations in the synthetic environment to encourage the model to learn robust features that transfer to real-world data (Tobin et al., 2017). Another class of methods focuses on domain adaptation by aligning feature distributions between synthetic and real domains through ad-

versarial learning or feature-level constraints (Ganin & Lempitsky, 2016; Tzeng, Hoffman, Saenko, & Darrell, 2017). While these methods can improve generalization, they often require large synthetic datasets or substantial amounts of real-world data.

In automated industrial inspection scenarios, such as surface defect detection, assembly verification, and quality control on production lines, data acquisition is challenging. It may require halting production, setting up specialized imaging systems, or capturing rare defect cases that occur infrequently. Moreover, defects are typically imbalanced and diverse, making it difficult to obtain sufficient labeled samples that cover all failure modes. Variations in materials, illumination conditions (e.g., reflections from metallic surfaces), and sensor configurations further increase the complexity of real-world data collection. As a result, building large-scale, fully annotated datasets for industrial inspection is both costly and operationally disruptive. Therefore, developing data-efficient sim-to-real adaptation methods remains an important research challenge. Instead of relying on large real datasets, it is desirable to identify a small number of representative real samples that provide maximum coverage of the synthetic feature space.

In this work, we propose a lightweight sim-to-real adaptation framework that combines representative sample selection and embedding-level alignment. First, we train an RF-DETR (Roboflow Detection Transformer) (Robinson, Robicheaux, Popov, Ramanan, & Peri, 2025) object detector using a synthetic dataset generated from a simulated environment. Although experimental results show that a model trained on synthetic data achieves near-perfect accuracy on the synthetic test set, its performance drops significantly when evaluated on real-world images, demonstrating a substantial domain gap.

To efficiently bridge this gap, we apply k-determinantal point process (k-DPP) sampling to select a small subset of diverse synthetic images based on their feature embeddings (Kulesza & Taskar, 2012). Corresponding real-world images are then collected for these selected samples, minimizing the amount of real data required for adaptation. To further reduce domain discrepancy, we introduce a feature-level embedding alignment strategy using Kullback–Leibler (KL) divergence. Specifically, the embeddings produced by the final decoder layer of the detector are aligned between paired synthetic and real images. This alignment encourages the model to produce similar high-level representations between real and synthetic domains, improving its ability to recognize real-world objects.

Experimental results demonstrate that the proposed approach significantly improves real-world detection performance using only a small number of real images. The combination of representative sample selection and feature-space alignment provides a practical and scalable solution for sim-to-real transfer in real-world deployment scenarios.

The main contributions of this work are summarized as follows:

- We propose a data-efficient sim-to-real adaptation framework for automated industrial inspection using synthetic data from a digital twin and a small set of paired real images, where the paired images are not necessarily visually identical.
- We introduce an active sample acquisition strategy based on k-DPP to select representative synthetic samples for real-world data collection.
- We propose a KL-divergence-based embedding alignment method that reduces the feature-space discrepancy between paired synthetic and real images.
- We demonstrate through ablation studies that the combination of k-DPP sample selection and KL alignment achieves the best real-world detection performance while requiring limited paired real images.

2. RELATED WORK

2.1. Object Detection

Object detection has evolved rapidly with the development of deep convolutional neural networks and large annotated datasets. Modern detectors can generally be categorized into single-stage and two-stage architectures.

Single-stage detectors prioritize efficiency and real-time performance. The YOLO (You Only Look Once) family of detectors formulates object detection as a single regression problem that directly predicts bounding boxes and class probabilities from full images (Redmon, Divvala, Girshick, & Farhadi, 2016). With the rapid evolution of Ultralytics releases, the latest YOLO26 (Jocher & Qiu, 2026) significantly improves both detection accuracy and computational efficiency. Owing to their speed and simplicity, YOLO-based models have been widely adopted in industrial and real-time applications, including manufacturing inspection(Chou, Wang, & Mao, 2025; Zuo, Dong, Gao, & Wu, 2024) and robotics(Zhao et al., 2026; Wu, Chen, Yu, & Li, 2026).

Transformer-based detectors represent another major line of development. DETR (DEtection TRansformer) (Carion et al., 2020) introduces an end-to-end object detection framework that formulates detection as a set prediction problem using a transformer encoder-decoder architecture. Unlike traditional detectors, DETR eliminates the need for hand-designed components such as anchor boxes and non-maximum suppression. Subsequent works such as Deformable DETR address the slow convergence and multi-scale limitations of the original model by introducing deformable attention mechanisms (Zhu et al., 2021; Robinson et al., 2025). Transformer-based detectors provide a flexible architecture for feature learning

and have become increasingly popular in research and industrial applications.

2.2. Sim-to-Real Transfer

One common approach to address sim-to-real gap is domain randomization, which introduces significant variability in synthetic environments to encourage models to learn domain-invariant representations (Tobin et al., 2017). Synthetic datasets have also been explored in industrial inspection and aerial detection applications where collecting real data is expensive (Wu, Guo, Tan, et al., 2024). Despite these efforts, the performance gap between synthetic and real domains remains a major challenge (Ruter, Durak, & Dauer, 2024).

2.3. Domain Adaptation and Feature Alignment

Domain adaptation techniques aim to reduce distribution differences between training and deployment environments by aligning feature representations across domains. Adversarial domain adaptation methods encourage the learning of domain-invariant features through adversarial training (Ganin & Lempitsky, 2016; Tzeng et al., 2017). More recent work explores feature-level alignment strategies that progressively reduce domain discrepancies between synthetic and real data (Gao, Wang, Yang, & Wu, 2025).

2.4. Representative Sample Selection

Selecting informative and diverse samples is a key strategy for improving data efficiency in both active learning and domain adaptation settings. Active learning methods aim to reduce annotation cost by identifying the most informative samples for labeling, with representative approaches including core-set selection (Sener & Savarese, 2018), BADGE (Ash, Zhang, Krishnamurthy, Langford, & Agarwal, 2020), and active learning literature (Settles, 2009). These methods typically rely on uncertainty estimation, gradient information, or feature diversity to construct compact training subsets that preserve performance under limited labeling budgets.

In contrast to uncertainty-based sampling, Determinantal Point Processes (DPPs) provide a probabilistic framework for directly modeling diversity in a dataset by selecting subsets that maximize coverage of the underlying feature space (Kulesza & Taskar, 2012). The k-DPP variant further enables selection of a fixed-size subset, making it particularly suitable for dataset summarization and budget-constrained selection.

3. METHODOLOGY

3.1. Digital Twin and Synthetic Data Generation

We first construct a simulated experimental setup to generate synthetic training data with NVIDIA Isaac Sim, shown in Figure 1. The scene replicates a typical room in an industrial setting. On one side of the wall, a pressure gauge and an electric box are mounted. Dome lighting in the simulation, which is a type of environment light that emits uniform illumination, is used to mimic the ceiling light in the real room. It helps produce realistic lighting and soft shadows in the synthetic images. A camera is initialized within the virtual environment, and its position, orientation (yaw, pitch, roll), and zoom scale are randomized via script. Each captured image is saved along with its camera parameters and corresponding object annotations.

Using this setup, a total of 1100 synthetic images were generated. Among them, we use 550 synthetic images for training the RF-DETR detector. A separate set of 550 synthetic images is used for validation and synthetic-domain evaluation, ensuring that model selection is performed without exposure to real-world data. A relatively large synthetic validation set was used because synthetic data can be generated at negligible cost, allowing a stricter evaluation of detector performance during model selection. All images may or may not contain the target pressure gauge, under varying viewpoints and imaging conditions.

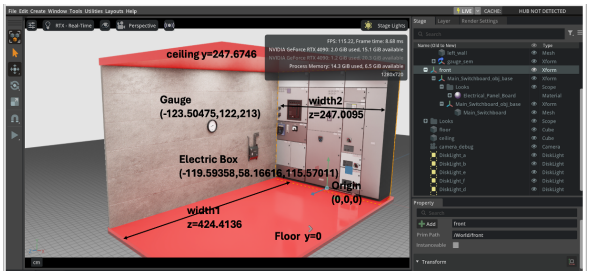

Figure 1. Digital twin setup for synthetic image generation.

To evaluate sim-to-real transfer, we then construct a corresponding real-world demonstration room, as shown in Figure 2. The room contains the same printed wall patterns as the synthetic environment. However, several unintentional discrepancies exist between the simulated and real settings. Unlike the synthetic scene, which contains a red floor and ceiling, the real environment uses a grey carpeted floor and a white ceiling partially covered by a net structure. Furthermore, the pressure gauge used in the real environment differs slightly in appearance from the synthetic model. These differences introduce realistic domain discrepancies in lighting, texture, color, and object appearance, creating a representative sim-to-real adaptation challenge.

For sim-to-real training, we collect 50 paired real-world images corresponding to synthetic samples selected by the k-DPP procedure described in Section 3.3. Although the synthetic and real images are not pixel-aligned and contain appearance differences, each pair is captured using approximately the same camera pose and viewing geometry. Therefore, the paired images share similar scene layout and object location, allowing feature alignment to focus on appearance-

Figure 2. Real World Setting.

related domain discrepancies rather than geometric variation.

To evaluate real-world performance, we additionally collect a separate held-out test set of 50 real images that are not used during training, embedding alignment, hyperparameter selection, or visualization. This separation ensures that all reported detection metrics reflect real-world generalization rather than memorization of the paired training samples.

The overall dataset composition and usage are summarized in Table 1.

Table 1. Dataset composition and split.

Set	Synthetic	Real
Train	550	0
Synthetic Validation	550	0
Adaptation	50(Subset of Train)	50
Test	0	50

3.2. Baseline Training on Synthetic Images

RF-DETR (Robinson et al., 2025) was chosen primarily for being the latest DETR variant, and similar results could likely be obtained with other DETR-based models. We finetune the model from their public checkpoint, solely on the 550 synthetic images. For all experiments, we freeze all layers until the last decoder layer to reduce training complexity, as the task involves a single, simple object, train on a single RTX 4090 GPU. The trained model achieves high performance on the synthetic validation set of 550 images. When evaluated on real images of the same printed gauge, detection works reliably. However, testing on a different type of real gauge demonstrates poor performance, with only a few instances detected, illustrating the sim-to-real gap (more details provided in the Results section).

3.3. Representative Synthetic Sample Selection via k-DPP

Unlike passive sim-to-real adaptation approaches that rely on randomly collected real-world data, our framework actively identifies informative samples for real-world acquisition through a k-Determinantal Point Process (k-DPP). Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be the set of N candidate synthetic images. For each image x_i, we extract a feature vector $\mathbf{f}_i \in \mathbb{R}^{256}$ by mean-pooling the 300 object-query embeddings from the final decoder layer of a pretrained RF-DETR model. We then construct the kernel matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ using the cosine similarity between ℓ_2-normalized feature vectors:

$$\mathbf{L}_{ij} = \frac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \, \|\mathbf{f}_j\|_2} \tag{1}$$

Since cosine similarities of ℓ_2-normalized vectors lie in $[-1, 1]$, and to ensure $\mathbf{L}$ is positive semidefinite, entries are shifted as $\mathbf{L}_{ij} \leftarrow \frac{1}{2}(\mathbf{L}_{ij} + 1)$, mapping values to $[0, 1]$. A k-DPP defines a probability distribution over subsets $Y \subseteq \mathcal{X}$ of size k:

$$\Pr(Y) = \frac{\det(\mathbf{L}_Y)}{\sum_{|S|=k} \det(\mathbf{L}_S)}, \quad |Y| = k \tag{2}$$

where $\mathbf{L}_Y$ is the principal submatrix of $\mathbf{L}$ indexed by Y. The determinant $\det(\mathbf{L}_Y)$ is geometrically interpreted as the squared volume of the parallelotope spanned by the feature vectors in Y: subsets whose embeddings are mutually dissimilar (i.e., span a larger volume) receive higher probability. By sampling subsets according to the k-DPP distribution, subsets with larger determinants are more likely to be selected. This minimises redundancy among selected images while preserving the main structural variation of the data, allowing a small representative subset to effectively approximate the full collection (Kulesza & Taskar, 2012).

For our experiments, k=50 images are selected.

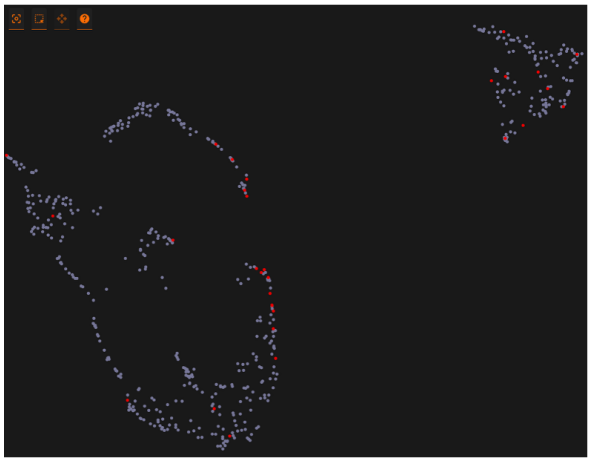

Figure 3. UMAP embedding in FiftyOne.

Using the FiftyOne (Moore & Corso, 2020) feature visualization technique, Figure 3 visualizes the UMAP (Uniform Manifold Approximation and Projection) embedding of all

550 synthetic images in grey dots, with the 50 selected samples highlighted in red. The corresponding 50 images are then collected from the real world using the recorded camera parameters from the synthetic counterparts, forming paired synthetic-real samples. Figure 4 shows an example pair of images, where the left image is the synthetic rendering and the right image is the corresponding real-world image captured using the recorded camera configuration. Note that the images are still not perfectly aligned at pixel level due to differences in lighting conditions and subtle variations in camera positioning.

Figure 4. Synthetic and Real Image Pair Example.

3.4. Sim-to-Real Training

We first construct the sim-to-real training dataset by pairing each selected synthetic image(50 images) with its corresponding real-world image. Subsequently, we conducted the training by applying the following two strategies:

Individual Training We established our baseline model using individual training, where real images were added as independent samples without any explicit alignment. The model was trained jointly on synthetic(550) and real images(50), totaling 600 images. All layers except the last decoder layer were frozen, and only the final decoder layer and detection head were fine-tuned following the standard RF-DETR object detection procedure (Robinson et al., 2025).

KL Divergence Alignment Next, we integrated our proposed embedding alignment method to reduce the sim-to-real gap. Specifically, we aim to bring the feature embeddings of paired synthetic and real images closer in the final decoder layer of RF-DETR. In our setup, we use 50 paired synthetic-real images for alignment, while the remaining 500 synthetic images are unpaired, resulting in a total of 600 training images, consistent with individual training.

In RF-DETR, the decoder outputs $N_q = 300$ object queries per image, each represented by a high-dimensional (256-D) embedding. Let $\mathbf{Q}^{\text{real}} = \{\mathbf{q}_1^{\text{real}}, \ldots, \mathbf{q}_{300}^{\text{real}}\}$ and $\mathbf{Q}^{\text{syn}} = \{\mathbf{q}_1^{\text{syn}}, \ldots, \mathbf{q}_{300}^{\text{syn}}\}$ denote the sets of 300 query embeddings for a real image and its paired synthetic image, respectively, where each $\mathbf{q}_i \in \mathbb{R}^{256}$.

Probabilistic Modeling via Diagonal Gaussians. To align the feature distributions of paired images, we model each image's query embeddings as samples from a multivariate Gaussian distribution with diagonal covariance. Specifically, for each image, we fit a diagonal Gaussian distribution using maximum likelihood estimation over its 300 query embeddings:

$$\mu_{\text{real}} = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{q}_i^{\text{real}}, \quad \sigma^2_{\text{real}} = \frac{1}{N_q}\sum_{i=1}^{N_q}(\mathbf{q}_i^{\text{real}} - \mu_{\text{real}})^2 + \epsilon \tag{3}$$

$$\mu_{\text{syn}} = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{q}_i^{\text{syn}}, \quad \sigma^2_{\text{syn}} = \frac{1}{N_q}\sum_{i=1}^{N_q}(\mathbf{q}_i^{\text{syn}} - \mu_{\text{syn}})^2 + \epsilon \tag{4}$$

where $\mu \in \mathbb{R}^{256}$ and $\sigma^2 \in \mathbb{R}^{256}$ are the mean and variance vectors (computed element-wise), and $\epsilon = 10^{-6}$ is added for numerical stability. This yields two diagonal Gaussian distributions: $P_{\text{real}} = \mathcal{N}(\mu_{\text{real}}, \text{diag}(\sigma^2_{\text{real}}))$ and $P_{\text{syn}} = \mathcal{N}(\mu_{\text{syn}}, \text{diag}(\sigma^2_{\text{syn}}))$. The diagonal covariance assumption reduces computational complexity while capturing per-dimension variance, making it suitable for high-dimensional embeddings.

KL Divergence Loss.

We minimize the Kullback–Leibler (KL) divergence between the Gaussian distributions fitted to the paired real and synthetic decoder embeddings:

$$\mathcal{L}_{\text{KL}} = \text{KL}(P_{\text{real}} \| P_{\text{syn}}) = \frac{1}{2}\sum_{d=1}^{256}\left[\log\frac{\sigma^2_{\text{syn},d}}{\sigma^2_{\text{real},d}} - 1 + \frac{\sigma^2_{\text{real},d}}{\sigma^2_{\text{syn},d}} + \frac{(\mu_{\text{real},d} - \mu_{\text{syn},d})^2}{\sigma^2_{\text{syn},d}}\right] \tag{5}$$

where d indexes the 256 embedding dimensions. This closed-form expression is derived from the KL divergence between two diagonal multivariate Gaussians.

We employ the forward KL divergence $\text{KL}(P_{\text{real}} \| P_{\text{syn}})$ to measure the discrepancy between the embedding distributions of paired real and synthetic images. Forward KL places a larger penalty on regions where the real-image embedding distribution has high probability but the synthetic distribution does not, encouraging the synthetic embeddings to better match feature patterns observed in real images. We adopt forward KL rather than symmetric alternatives such as Jensen–Shannon divergence because its asymmetric formulation aligns with the objective of adapting synthetic representations to better match real-world feature statistics.

This encourages paired synthetic-real samples to have similar high-level feature representations in the last decoder layer, promoting better generalization to real-world images. The KL loss, $\mathcal{L}_{\text{KL}}$, is added to the standard detection loss (bounding box regression and classification) with a weighting factor of λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{detection}} + \lambda \mathcal{L}_{\text{KL}} \tag{6}$$

During training, the 50 paired samples contribute both detection and alignment losses, while the 500 unpaired synthetic images contribute only to the detection loss, allowing the model to leverage the majority of synthetic data while learning representative features from the small set of real images.

KL Loss Weight. To study the sensitivity of the KL loss weighting factor, we evaluate different values of $\lambda \in \{1, 5, 10\}$. We observe that a small weight ($\lambda = 1$) leads to insufficient alignment between synthetic and real feature distributions, while a large weight ($\lambda = 10$) over-emphasizes the alignment term and slightly degrades detection performance due to reduced task-specific discrimination in the embedding space. The intermediate value $\lambda = 5$ achieves the best trade-off between domain alignment and detection accuracy, and is therefore used as the default setting in all experiments.

3.5. Ablation Study: Random Selection and L2 Alignment

To evaluate the effect of sample selection and embedding alignment strategy, we conducted a series of ablation studies along four key axes:

- **Choice of k in k-DPP:** We initially set k = 50, corresponding to approximately 10% of the synthetic training set, as a trade-off between annotation cost and coverage of the feature space. To evaluate the sensitivity of the proposed method to the choice of k, we further conduct experiments with k = 25 and k = 75, representing lower- and higher-budget sampling regimes, respectively. This analysis allows us to assess the robustness of k-DPP-based sample selection under different data acquisition budgets.
- **L2 Alignment:** Instead of using KL divergence to align synthetic and real embeddings, we use L2 distance. First, we compute the mean embedding across all $N_q = 300$ decoder queries for each image:

$$\bar{\mathbf{q}}^{\text{syn}} = \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbf{q}_i^{\text{syn}}, \quad \bar{\mathbf{q}}^{\text{real}} = \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbf{q}_i^{\text{real}} \tag{7}$$

  We then minimize the L2 distance between the averaged embeddings:

$$\mathcal{L}_{\text{L2}} = \left\| \bar{\mathbf{q}}^{\text{syn}} - \bar{\mathbf{q}}^{\text{real}} \right\|_2^2 \tag{8}$$

  This approach encourages the overall feature representation of each synthetic-real pair to be similar, while being simpler and computationally cheaper than full distribution alignment. It provides a baseline to assess the benefit of distribution-level KL divergence.
- **KL Alignment Plus L2 Maximization:** In addition to aligning paired synthetic and real embeddings, we push non-paired synthetic embeddings away to maintain latent space structure. Let $P_{\text{syn}}^{(i)}$ and $P_{\text{real}}^{(i)}$ denote the diagonal Gaussian distributions estimated from the decoder query embeddings of the i-th synthetic-real pair. We compute the corresponding mean embeddings $\mu_{\text{syn}}^{(i)}$ and $\mu_{\text{real}}^{(i)}$ as described previously.
  The combined objective is defined as:

$$\begin{aligned} \mathcal{L}_{\text{KL+L2}} = & \frac{1}{N} \sum_{i=1}^{N} D_{\text{KL}}\left(P_{\text{real}}^{(i)} \,\|\, P_{\text{syn}}^{(i)}\right) \\ & - \lambda \cdot \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} \left\| \mu_{\text{syn}}^{(j)} - \mu_{\text{real}}^{(i)} \right\|_2^2 \end{aligned} \tag{9}$$

  Here, the KL divergence term aligns paired synthetic-real embeddings, while the L2-based pushing term encourages unpaired synthetic embeddings to remain distributed across the latent space, preserving diversity. The balancing weight λ is set to 1 in our experiments for simplicity. This ablation allows us to study the benefit of preserving latent structure alongside pairwise alignment.
- **Random Selection:** To assess the impact of our k-DPP sample selection strategy, we also randomly select 50 synthetic images to pair with real images. Both KL divergence alignment and individual training are applied with these randomly paired samples. This experiment evaluates whether careful selection of diverse synthetic samples is necessary for effective adaptation or whether simple random pairing suffices.

For all experiments, the same hyperparameters are used, 50 training epochs with model selection based on the validation set, ensuring fair comparison across all methods.

4. RESULT AND DISCUSSION

4.1. Baseline Performance on Synthetic Images

We first trained the RF-DETR model on 550 synthetic images, as a single-category object detection task. Because the target object has a relatively simple geometry, the model achieves near-perfect performance on the synthetic validation set, with an mAP@0.5:0.95 of 0.982. These results indicate that RF-DETR is powerful to reliably detect the gauge and confirm that the model architecture and training procedure are sufficient for the controlled synthetic domain.

When evaluating the model on real-world images, we ob-

served two different outcomes:

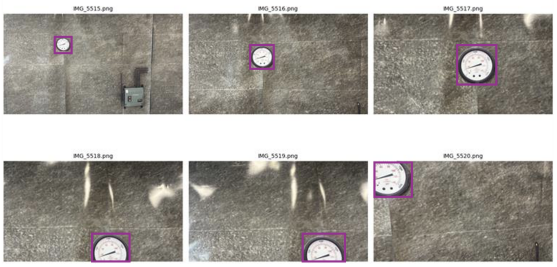

Figure 5. Detection Results on Exact Same Printed Gauge.

When testing on real-world images of the exact same printed gauge, Figure 5 shows the model successfully detected the object in the pink bounding boxes, confirming that the synthetic data provides sufficient signal for identical instances.

However, when evaluated on a visually different but geometrically similar gauge, detection failed entirely, highlighting the sim-to-real gap, shown in Figure 6. This indicates that although synthetic data can provide strong initial supervision, it cannot fully capture the variability present in real-world scenarios.

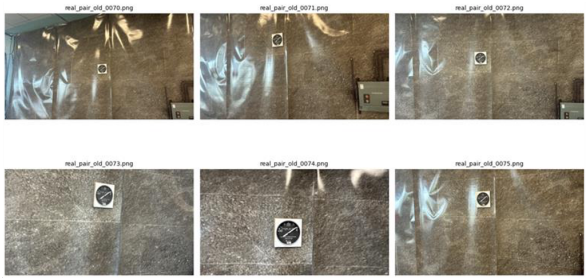

Figure 6. Detection Results on Slightly Different Gauge.

Note that the constructed digital twin does not perfectly replicate the real-world environment, which introduces some unavoidable discrepancies. For example, in the digital twin, the ceiling and floor are rendered in red, whereas in reality the floor is a gray carpet and the ceiling consists of a black net with white panels. Additionally, the figures reveal significant differences in color, lighting, and reflections between the synthetic and real scenes.

4.2. Embedding Visualization

To understand how domain differences manifest in the model's latent space, we visualized the embeddings from the last decoder layer using UMAP. We included the full set of 600 images, consisting of 500 unpaired images and 50 synthetic images paired with 50 real images. In the visualizations, orange dots represent the 500 unpaired synthetic images, red dots represent the 50 paired synthetic images, and grey dots correspond to the 50 paired real images.

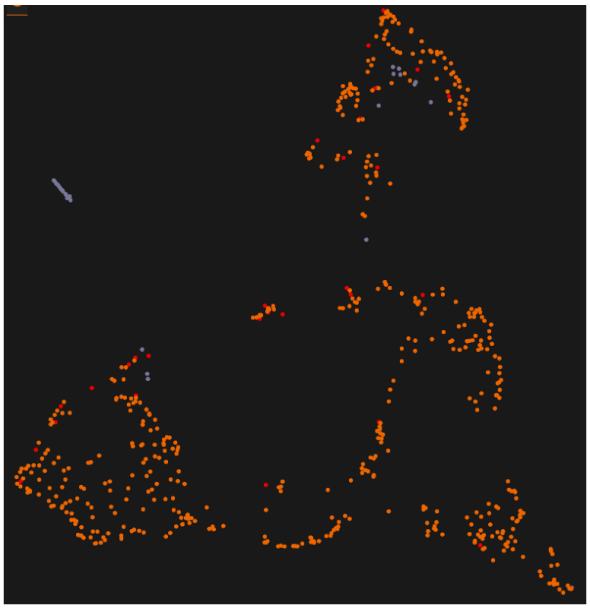

Figure 7. UMAP visualization of the last decoder layer embeddings after individual training.

Figure 7 shows the embeddings after individual training without any alignment. Here, the orange and red dots (synthetic images) are closely clustered together, while the grey dots (real images) form a separate cluster. This separation indicates that the model encodes real and synthetic features differently, with the unaligned latent space unable to bridge the sim-to-real gap.

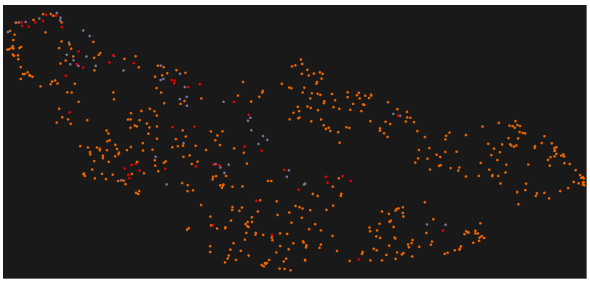

Figure 8. UMAP visualization of the last decoder layer embeddings after KL divergence alignment.

Figure 8 illustrates the effect of the KL divergence alignment. In this case, paired red and grey dots are drawn closer together, and the red synthetic dots are distributed within the broader orange synthetic cluster. This demonstrates that KL alignment effectively reduces the domain shift by bringing paired synthetic and real embeddings closer while retaining the overall synthetic distribution.

Beyond qualitative clustering, we quantified the embedding alignment using the average pairwise distance between synthetic-

real pairs before and after alignment. KL-aligned embeddings reduced the average distance by approximately 31.4%, compared to individual training, confirming that feature-level alignment is effective in bringing the domains closer. Such visualization and quantitative assessment provide evidence that improvements in downstream detection metrics are not only due to the inclusion of real images but are also mediated by enhanced latent space coherence.

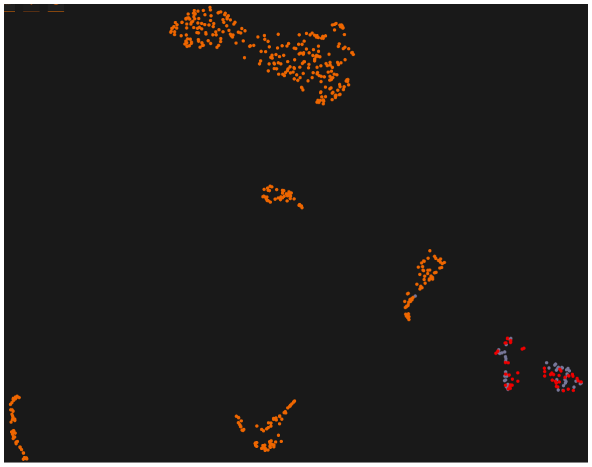

Figure 9. UMAP visualization of the last decoder layer embeddings after KL alignment plus L2 Maximization.

Finally, Figure 9 shows embeddings under KL alignment with L2 distance maximization for non-paired samples. Paired red and grey dots remain tightly aligned, while unpaired synthetic images (orange) form separate clusters. This L2 maximization explicitly separates non-paired images, improving the distinction between paired and unpaired samples. While paired images are encouraged to stay close, we also ensure that each paired synthetic image does not deviate significantly from the overall synthetic image distribution.

Quantitatively, the average pairwise distance between paired embeddings decreases significantly from individual training to KL alignment and KL alignment with pushing, confirming that closer latent representations correlate with improved detection performance. These visualizations highlight that alignment strategies not only bring paired images together but can also structure the latent space to better separate informative pairs from unrelated samples, which is critical for reducing the sim-to-real gap.

Since the original embeddings are 256-dimensional, UMAP projects them into a two-dimensional space for visualization. This dimensionality reduction inevitably compresses part of the information contained in the embeddings. Therefore, the visualization is intended only as a qualitative reference rather than a precise representation of the embedding distribution.

4.3. Quantitative Evaluation

Table 2 summarizes detection performance on 50 real-world images across different training strategies, including mAP@[0.50:0.95], confusion matrices, and F1 score.

Several key observations can be drawn from these results:

- **Confusion matrix analysis:** Across all evaluated methods, the number of false positives remains consistently zero, while performance differences are primarily driven by variations in false negatives. This behavior is expected in our experimental setting, which is a simple single-class detection task where the pressure gauge has a strong visual distinction from the background. As a result, the detector rarely confuses background regions as the target object, and no background regions are predicted as false positives across all methods. Therefore, improvements from the proposed method are primarily reflected in reduced false negatives (i.e., improved recall), rather than changes in precision.
- **Baseline synthetic-only performance:** Training solely on synthetic images yields very low mAP (0.2516) and F1 score (0.434) on real images, despite perfect coverage of synthetic samples. This confirms that without real-world data, the model fails to generalize, highlighting the significant sim-to-real gap.
- **Effect of curated real samples (k-DPP selection):** Incorporating representative real images selected via k-DPP dramatically improves detection performance. Individual training on these 50 paired real images increases mAP to 0.7421 and F1 to 0.867. This suggests that selecting diverse representative samples provides more useful adaptation data than random acquisition under the same labeling budget.
- **Impact of embedding alignment (KL alignment):** Applying KL divergence to align paired synthetic and real embeddings further enhances performance, achieving mAP (0.8853) and F1 (0.944). This indicates that reducing the latent-space distance between paired embeddings allows the model to generalize better to real images while retaining synthetic knowledge.
  To further evaluate the robustness of the results under the relatively small real test set, we conduct 10 independent runs with random initializations. Each training run can be completed within approximately one hour on a single RTX 4090 GPU, without requiring multi-GPU training or additional pretraining stages beyond the frozen backbone. The proposed method consistently achieves an average mAP of 0.8913 ± 0.041 and F1 score of 0.956 ± 0.023, demonstrating stable performance and low variance across runs, while maintaining low computational overhead suitable for practical industrial deployment.
- **Effect of k in k-DPP sampling:** We analyze the sensitivity of the proposed method to different values of k.

Table 2. Detection performance on 50 real-world test images for different training strategies.

Method	mAP@50:95	Confusion Matrix [TP, FN, FP, TN]	F1
Synthetic Only	0.2516	[13,34,0,3]	0.434
k-DPP + Individual Training	0.7421	[36,11,0,3]	0.867
k-DPP + KL Alignment	0.8853	[42,5,0,3]	0.944
k-DPP(k=25) + KL Alignment	0.7534	[38,9,0,3]	0.895
k-DPP(k=75) + KL Alignment	0.9122	[44,3,0,3]	0.967
k-DPP + L2 Alignment	0.7315	[35,12,0,3]	0.854
k-DPP + KL Alignment + L2 Maximization	0.6933	[33,14,0,3]	0.825
Random Selection + Individual Training	0.7061	[33,14,0,3]	0.825
Random Selection + KL Alignment	0.7945	[38,9,0,3]	0.894

When reducing k to 25, the performance drops to 0.7534 mAP and 0.895 F1, indicating insufficient coverage of the synthetic feature space. Increasing k to 75 improves performance to 0.9122 mAP and 0.967 F1, slightly outperforming the default setting of k=50 (0.8853 mAP). Although k=75 achieves the highest performance, it requires 50% more paired real images than k=50.
While there is no strict theoretical rule for selecting an optimal k, it is primarily determined by a trade-off between annotation cost and coverage of the synthetic feature space in practical industrial inspection settings. In all main experiments, we use k=50 as a consistent default setting, as it provides a balanced performance-cost trade-off.

- **KL alignment plus L2 maximization:** This variant achieves slightly lower mAP (0.6933) and F1 (0.825) compared to KL alignment alone. While maximizing the distance between non-paired embeddings can improve latent-space structure, in this small dataset it causes the paired synthetic images to deviate too far from the overall synthetic distribution, slightly limiting detection performance on real images. It suggests that applying a smaller weight (λ) to the maximization term could reduce over-separation and help the synthetic images remain within the same distribution, potentially yielding better results.
- **L2 alignment versus KL alignment:** Using L2 alignment achieves mAP 0.7315 and F1 0.854, slightly worse but similar to Individual Training (mAP 0.7421, F1 0.867). This is likely because the L2 distance is computed by first taking the mean of the 300 queries in each dimension. Averaging in this way is somewhat arbitrary and fails to capture the full distribution of the query embeddings, which can reduce the effectiveness of alignment. In contrast, KL divergence considers the full distribution, leading to better alignment and higher detection performance.
- **Random selection strategies:** Both random selection with individual training and KL alignment improve over the synthetic-only baseline but generally underperform k-DPP selection. It confirms that intelligent selection of real samples is critical for maximizing sim-to-real transfer efficiency.

These results collectively demonstrate that a combination of careful sample selection and latent-space alignment is the most effective approach for reducing the sim-to-real gap. Even with only 50 real-world images, the model achieves substantial improvements in both mAP and F1, highlighting the efficiency of the proposed sim-to-real transfer strategies.

5. CONCLUSION

In this work, we presented a data-efficient sim-to-real adaptation framework for automated industrial inspection that combines representative sample selection through k-DPP with feature-level alignment using KL divergence. The proposed approach leverages a digital twin to generate synthetic training data and requires only a small number of paired real images for adaptation.

- Training on synthetic images alone provides high performance on synthetic validation but fails to generalize to real-world variations, highlighting the sim-to-real gap.
- Selecting representative real images via k-DPP significantly improves real-world detection, even with only 50 paired samples.
- Embedding alignment, especially KL divergence, further reduces the sim-to-real discrepancy, achieving the highest mAP and F1 scores. L2 alignment performs slightly worse due to its simplistic mean-based distance calculation, which fails to fully capture the query distribution.
- Pushing non-paired embeddings can structure the latent space but may reduce performance if over-applied, indicating that hyperparameter tuning (e.g., λ) is important.

The primary novelty of this work lies in integrating active sample acquisition and feature alignment into a unified data-efficient sim-to-real adaptation framework.

Limitations Despite the promising results, several limitations should be acknowledged:

- The current evaluation is conducted on a single object category within a controlled inspection environment. Additional studies are required to validate generalization to multiple object classes and more diverse industrial settings.
- The real-world dataset remains relatively small due to the cost of collecting paired synthetic-real samples, although repeated runs demonstrate stable performance.
- The selection of k and the KL-loss weighting factor are empirically determined. While sensitivity studies show the method is reasonably robust, a principled strategy for selecting these hyperparameters remains an open question.
- The digital twin does not perfectly replicate the real environment, and the effectiveness of the proposed framework under larger appearance and environmental discrepancies requires further investigation.

Future Work Several directions may further improve the proposed framework:

- **Adaptive sample selection:** Develop principled criteria for automatically determining the optimal value of k based on dataset coverage, diversity, or annotation budget.
- **Alignment objective design:** Explore alternative distribution-alignment objectives, including symmetric KL divergence, Jensen–Shannon divergence, and contrastive representation learning approaches.
- **Multi-object industrial inspection:** Extend the framework to multiple object categories, defect types, and cluttered inspection environments to evaluate generalization beyond the current proof-of-concept setting.
- **Broader industrial validation:** Apply the approach to more complicated industrial inspection tasks, including defect detection, assembly verification, and quality-control applications involving different sensors and lighting conditions.
- **Active sim-to-real adaptation:** Investigate adaptive data acquisition strategies that iteratively select new real samples based on model uncertainty or feature-space coverage, further reducing data collection costs.

These directions will help make sim-to-real detection more efficient, adaptive, and broadly applicable to diverse real-world industrial inspection tasks.

REFERENCES

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2020). Deep batch active learning by diverse, uncertain gradient lower bounds. In *International conference on learning representations (iclr)*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision (eccv)*.

Chou, P.-H., Wang, C.-C., & Mao, W.-L. (2025). Yolo-based defect detection for metal sheets. *arXiv preprint arXiv:2509.25659*.

Ganin, Y., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Gao, D., Wang, Q., Yang, J., & Wu, J. (2025). Domain adaptive object detection via synthetically generated intermediate domain and progressive feature alignment. *Image and Vision Computing*, *154*, 105404.

Jocher, G., & Qiu, J. (2026). *Ultralytics yolo26*. Retrieved from `https://github.com/ultralytics/ultralytics`

Kulesza, A., & Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, *5*(2–3), 123–286.

Moore, B. E., & Corso, J. J. (2020). Fiftyone. *GitHub. Note: https://github.com/voxel51/fiftyone*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Ieee conference on computer vision and pattern recognition (cvpr)*.

Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., & Peri, N. (2025). Rf-detr: neural architecture search for real-time detection transformers. *arXiv preprint arXiv:2511.09554*.

Ruter, J., Durak, U., & Dauer, J. (2024). Investigating the sim-to-real generalizability of deep learning object detection models. *Journal of Imaging*.

Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *International conference on learning representations (iclr)*.

Settles, B. (2009). Active learning literature survey. *University of Wisconsin-Madison*.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Cvpr*.

Wu, Y., Chen, J., Yu, X., & Li, J. (2026). Yolo-foa: A lightweight rotational target detection algorithm based on improved yolo for optical fiber robot. *Biomimetic Intelligence and Robotics*, 100273.

Wu, Y., Guo, W., Tan, Z., et al. (2024). Syn2real detection in the sky: Generation and adaptation of synthetic aerial

ship images. *Applied Sciences*.

Zhao, H., Guo, J., Dong, E., Guo, R., Zhao, L., Wang, C., ... Li, Y. (2026). Yolo-gdcnn: Real-time operating point detection for live working robots in the power industry. *High Voltage*.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. In *Iclr*.

Zuo, Z., Dong, J., Gao, Y., & Wu, Z. (2024). Hyperdefect-yolo: Enhance yolo with hypergraph computation for industrial defect detection. *arXiv preprint arXiv:2412.03969*.