University of Kent

Kent Academic Repository

Yu, Jongmin, Oh, Hyeontaek, Sun, Zhongtian, Harit, Anoushka, Yoo, Namhyun and Yang, Jinhong (2026) *Prototype-guided diffusion model for multi-class unsupervised anomaly detection*. IEEE Access, 14 . pp. 142986-142998.

Downloaded from
https://kar.kent.ac.uk/116728/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.1109/ACCESS.2026.3732635

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 August 2026, accepted 1 September 2026, date of publication 10 September 2026, date of current version 21 September 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3732635

RESEARCH ARTICLE

Prototype-Guided Diffusion Model for Multi-Class Unsupervised Anomaly Detection

JONGMIN YU[1], HYEONTAEK OH[1], ZHONGTIAN SUN[2], ANOUSHKA HARIT[3], NAMHYUN YOO[4], (Member, IEEE), AND JINHONG YANG[5], (Member, IEEE)

[1]AI Laboratory, nextstudio Inc., Daejeon 34130, Republic of Korea
[2]School of Computing, University of Kent, Kent, CT2 7NZ Canterbury, U.K.
[3]Cancer Research, University of Cambridge, B2 1TN Cambridge, U.K.
[4]School of Computer Engineering and Science, Kyungnam University, Changwon-si 51767, Republic of Korea
[5]Department of Medical Information Technology, Inje University, Gimhae-si 50834, Republic of Korea

Corresponding author: Jinhong Yang (jinhong@inje.ac.kr)

This work was supported by the National IT Industry Promotion Agency (NIPA) grant funded by Korean Government (MSIT) (Development of Adaptive Physics-Aware Synthetic Data Generation Technology) under Grant RS-2026-25621689.

ABSTRACT Unsupervised Anomaly detection is important in industrial inspection and automation, where defects are rare, stochastic, and costly to annotate, while nominal data are abundant. Diffusion models have shown strong potential for unsupervised anomaly detection, where only normal data are available for training. However, standard class-conditional diffusion models rely on a single embedding per class, which insufficiently captures the multi-modal nature of normal distributions in unified multi-class settings. This coarse conditioning can yield over-generalised reconstructions that obscure anomalies or fail to preserve fine-grained normal structures. We propose ProtoDiffAD, a prototype-guided diffusion framework that represents each class using a dictionary of normal prototypes learned via a pretrained variational autoencoder. We further introduce Multi-Prototype Guidance (MPG), which dynamically aggregates relevant prototypes through attention during denoising. This mechanism enforces reconstruction fidelity to the normal manifold, enhancing anomaly localisation. Experiments on standard benchmarks (MVTec-AD and VisA) demonstrate consistent improvements over existing unified methods. The source code will be publicly available.

INDEX TERMS Prototype-guided, diffusion model, unsupervised anomaly detection (UAD).

I. INTRODUCTION

Unsupervised Anomaly Detection (UAD) is a fundamental computer vision task and has great potential for many industrial areas such as autonomous maintenance [1], robotics [2] and quality management [3], [4]. It aims to identify samples that deviate from a normal distribution when only normal data are available for training. This paradigm is particularly important in industrial inspection, where defects are rare, stochastic, and costly to annotate, while nominal data are abundant. Traditional methods typically follow a *"one-class-one-model"* paradigm, training a separate model for each category. Recent research has shifted towards unified multi-class anomaly detection, where a single model detects anomalies across multiple categories simultaneously. Although this unified framework improves scalability and deployment flexibility, it introduces the challenge of modelling substantial inter-class variability while preserving sensitivity to subtle intra-class deviations.

Reconstruction-based generative models, such as Autoencoders (AEs) and Generative Adversarial Networks (GANs), have historically served as the backbone for UAD. These methods operate on the premise that a model trained exclusively on normal data will incur high reconstruction errors when processing anomalous input. However, these architectures often suffer from the "identity shortcut" [5] phenomenon, where the model generalises sufficiently to reconstruct anomalies alongside normal patterns, thereby diminishing detection performance.

Recently, diffusion probabilistic models have emerged as a potent alternative, exhibiting superior mode coverage and training stability compared to GANs. In the context

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

of unified multi-class UAD [3], [5], [6], these models typically rely on class-conditional generation, where a scalar class embedding [5] or context embedding [6] extracted from class labels guides the denoising trajectory. However, conditioning on a coarse class embedding is often suboptimal for anomaly detection. A single static embedding compresses the inherently multi-modal distribution of a category, which encompasses variations in pose, texture, and subtypes, into a single point in the latent space. This representation may be insufficient when the nominal distribution is multi-modal, encompassing substantial variations in viewpoint, pose, texture, geometry, illumination, or product subtype.

Consequently, the diffusion process may lack sufficiently fine-grained guidance to faithfully reconstruct instance-specific normal details. The model may produce an over-generalised reconstruction that either preserves an anomalous structure as a plausible within-class variation or modifies normal details that are not adequately represented by the category-level condition. This limitation can result in generic reconstructions that either hallucinate anomalous content as plausible class variations or fail to preserve the structural consistency of the input.

To challenge this issue, we propose the **Proto**type-guided **Diff**usion Model for **A**nomaly **D**etection (**ProtoDiffAD**). We reformulate the diffusion process using explicit, fine-grained normality guidance. Departing from the reliance on broad class labels, ProtoDiffAD anchors the generative process to a prototype dictionary constructed from the latent features using prepared normal datasets and a pre-trained Variational Autoencoder (VAE). This dictionary explicitly encapsulates the diverse 'normal modes' inherent within each category.

ProtoDiffAD is trained with the proposed Multi-Prototype Guidance (MPG). During the reverse diffusion process, rather than conditioning on a static label, the model attends to a dynamic mixture of prototypes that are contextually relevant to the current inputs. This mechanism effectively constrains the denoising trajectory, forcing the reconstructed image to reside near a learned normal prototype. Since anomalous regions fall outside this manifold, reproducing them becomes challenging for the model, resulting in pronounced discrepancies in the residual map, which facilitate precise anomaly localisation.

To show the effectiveness of the proposed method, we have evaluated ProtoDiffAD using MVTec-AD [7] and VisA [8] datasets, which are widely recognised industrial benchmarks. ProtoDiffAD produces an image-level AUROC of 99.5 and a pixel-level AUROC of 98.6 on the MVTec-AD dataset. It achieves an image-level AUROC of 97.1 and a pixel-level AUPRO of 93.8 on the VisA dataset. On the image-level AUROC, ProtoDiffAD achieves state-of-the-art (SOTA) performance, and on other pixel-level evaluations, it shows competitive performance and partially outperforms existing SOTA methods. Our ablation studies demonstrate that prototype-based guidance yields substantial improvements in both detection accuracy and localisation precision.

The main contributions of our paper are summarised as follows:

- **Prototype-Guided Diffusion Framework:** We present ProtoDiffAD, a unified multi-class unsupervised anomaly detection (UAD) architecture. The proposed framework replaces coarse class embeddings with a dictionary of fine-grained latent prototypes, thereby explicitly modelling the multi-modal nature of normality across diverse categories.
- **Multi-Prototype Guidance (MPG):** We introduce a reformulated guidance mechanism that dynamically aggregates relevant prototypes through an attention-based strategy. This design enables targeted corrections during the diffusion process, facilitating robust reconstruction of normal regions while suppressing anomalous features.
- **Superior Performance:** Extensive experiments on the MVTec-AD and VisA benchmarks demonstrate that ProtoDiffAD outperforms existing unified methods, achieving significant improvements in both image-level detection and pixel-level localisation.

II. RELATED WORK

Unsupervised anomaly detection (UAD) identifies observations that deviate from a nominal distribution using only normal training samples. Existing methods broadly comprise reconstruction-, embedding-, and anomaly-synthesis-based approaches. Reconstruction methods employ autoencoders, variational autoencoders, or generative adversarial networks and use reconstruction discrepancies as anomaly evidence [9], [10], [11], [12], [13], [14], [15]. Although their residual maps are interpretable, expressive models may reconstruct anomalous structures through the ''identity shortcut'' [5], while pixel residuals remain sensitive to benign variations in texture, illumination, and alignment. Embedding methods such as PaDiM [16] and PatchCore [4] detected feature-space deviations using distribution modelling or nearest-neighbour retrieval. Their performance, however, depended on backbone transferability and may incur substantial memory costs. Anomaly-synthesis methods, including CutPaste [17] and DRAEM [18], generated pseudo-defects for discriminative supervision, but their generalisation was constrained by the correspondence between synthetic and real anomalies [19], [20].

Conventional UAD commonly follows a *''one-class–one-model''* protocol. When the protocol is extended to the multi-class setting, computational and maintenance costs increase with the number of categories. Unified multi-class UAD instead trains a single detector across categories [3], [5], requiring it to accommodate inter-class variation whilst preserving sensitivity to subtle intra-class defects. UniAD [5] introduced neighbour-masked attention and feature jittering to reduce direct copying, whereas RD4AD [3] employed reverse distillation from a fixed teacher. Other methods used multimodal fusion, multi-scale reconstruction, semantic prompts, vision–language priors, memory mechanisms,

contrastive learning, or residual synthesis [1], [19], [20], [21], [22], [23], [24], [25], [26], [27]. Nevertheless, reconstruction constraints may suppress fine nominal details, teacher–student models inherit teacher biases, and global prompts or embeddings provide limited information about instance-specific pose, texture, viewpoint, or subtype.

Diffusion models have recently been adopted for anomaly reconstruction because their iterative reverse process permits normality guidance at multiple noise levels. DiAD [28] introduced diffusion-based unified anomaly detection, GLAD [29] combined global and local adaptive diffusion, and DeCo-Diff [6] reformulated denoising as correction towards normality. Their effectiveness nevertheless depended strongly on the conditioning representation. A fixed class or context embedding is shared across all observations of a category and generally remains unchanged during denoising. It may therefore identify the category without specifying the relevant normal configuration, leading to averaged reconstructions, loss of valid details, or preservation of anomalies as plausible intra-class variation.

Prototype-based learning offers a structured alternative to global embeddings. Prototypical Networks [30] represented each class through an embedding-space prototype, while SwAV [31] and prototypical contrastive learning [32] used prototypes to capture cluster-level organisation. However, a single prototype implicitly assumes a compact or approximately unimodal distribution and may inadequately represent variations in geometry, texture, pose, or viewpoint. Static prototypes also do not specify how anomalous observations should be restored towards normality. Prototype-like representations have also been applied to anomaly detection. MemAE [33] reconstructed inputs through sparse retrieval from a learnable memory of nominal patterns, whereas PatchCore [4] retained representative patch features for distance-based scoring. MemAE performed retrieval within a deterministic reconstruction pathway, while PatchCore does not generate a corrected image. Both additionally face a trade-off between normal-mode coverage and dictionary complexity. VPDM [34] more directly connected prototype learning with diffusion-based UAD. It constructs vague prototypes that preserve fundamental information whilst suppressing anomalous details, then applies diffusion to restore image content. Its objective is primarily to produce an information-restricted restoration condition; it does not model each category using multiple latent normal modes whose contributions are adaptively recomputed throughout denoising.

The proposed ProtoDiffAD combines unified diffusion reconstruction with multi-prototype normality modelling. It constructs a compact class-specific dictionary of normal latent prototypes and employs Multi-Prototype Guidance to compute soft relevance weights from the current noisy latent state. The prototype context is updated during denoising, allowing different observations within a category to be guided towards different combinations of normal modes. ProtoDiffAD therefore differs from single-prototype metric-learning methods, memory-bank approaches based on direct retrieval or distance scoring, and diffusion frameworks using static category-level conditions.

III. ProtoDiffAD

A. PRELIMINARIES

Denoising Diffusion Probabilistic Models (DDPMs) [35] are based on the idea of progressively perturbing data samples with noise via a forward process and then reversing it to generate new samples.

Let $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$ denote the data samples (*i.e.,* images), where $x^{(i)} \in \mathbb{R}^{H\times W\times C}$ is the i-th image, with H, W, and C representing its height, width, and number of channels, respectively. *Forward diffusion process* consists in gradually corrupting data samples by adding Gaussian noise for t timesteps where $t \in [1, T]$. Following [35], the forward noising process can be characterised as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

Here, β_t is a noise scheduling parameter that controls the amount of noise added at each step, such that $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for a sufficiently large T. The marginal distribution at the moment t can then be explicitly defined from x_0 as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{T}\alpha_i$ and $\alpha_t = 1-\beta_t$.

In the *reverse process*, which iteratively removes noise through a series of denoising steps to generate new samples from Gaussian noise, the prior probability at each step can be modeled as a Gaussian distribution:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_t\mathbf{I}), \tag{3}$$

in which β_t is fixed for each t, and the mean functions $\mu_\theta(x_t, t)$ parameterized by θ are trainable:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big). \tag{4}$$

This reverse process is trained with the variational lower bound of the log-likelihood of x_0:

$$\begin{aligned} l_\theta, = &-\log p_\theta(x_0|x_1) \\ &+ \sum_t \mathcal{D}_{KL}\big(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)\big). \end{aligned} \tag{5}$$

By applying the reparameterisation trick [9], the model can be trained using a simple mean-squared error loss between the ground truth sampled Gaussian noise and the predicted noise:

$$\mathcal{L} = \min_\theta \left(\mathbb{E}_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)\|_2^2\right], x\right) \tag{6}$$

with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, starting from a noise vector $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, sampling is achieved by iteratively applying the reverse process to gradually reconstruct a clean sample x_0.

In this work, we use Latent Diffusion Model (LDM) [36]. LDM conducts these processes in the latent feature space so that x in Eq. (1) and Eq. (3) will be replaced to a latent feature z, and the objective function (Eq. (6))

is reformulated with respect to z; thereby significantly reducing model size and computational complexity, which are critical for deploying diffusion-based systems in small-scale resource-constrained environments in robotics and industries.

B. METHODOLOGY OVERVIEW

We propose **ProtoDiffAD**, a unified framework for unsupervised anomaly detection that leverages the generative capability of diffusion models guided by fine-grained normality prototypes. Our core insight is that a single class embedding extracted from an explicit class label is insufficient to capture the intra-class variance of complex objects in a multi-class setting. Instead, we explicitly model the normal data manifold using a dictionary of latent prototypes.

The training pipeline consists of two distinct phases. In the first phase, we construct a **Prototype Dictionary** by mapping normal training samples into a latent space and identifying representative centroids via clustering. This dictionary encapsulates the multi-modal distribution of normality for each object category.

In the second phase, we train a conditional diffusion model to reconstruct normal samples. Unlike standard class-conditional models [6], [37], ProtoDiffAD employs a **Multi-Prototype Guidance (MPG)**. During the reverse diffusion process, the model is conditioned on a dynamic mixture of prototypes specific to the object class. By guiding the generation process toward the nearest "normal" modes defined by the prototypes, we ensure that anomalous regions, which lack corresponding prototypes, are reconstructed with high error, thereby facilitating detection.

C. PROTOTYPE-GUIDED DIFFUSION MODEL

1) PROTOTYPE DICTIONARY CONSTRUCTION

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denotes N numbers of multi-class normal samples, where x_i and y_i denote the i^{th} sample and class label, respectively. We extract latent features $Z_c = \{\mathcal{E}_\phi(x) \mid (x, y) \in \mathcal{D}, y = c\}$ for each class. To encapsulate the diverse "normal modes" (*e.g.,* different viewpoints), we apply K-Means clustering to Z_c. The prototypes are defined as the closest latent features from the centroids of these clusters, as follows:

$$\mathcal{P} = \{\mu_{c,k}\}_{c=1,k=1}^{C_{\text{cls}},K}, \tag{7}$$

where C_{cls} defines the total number of classes, and $\mu_{c,k} \in \mathbb{R}^D$ indicates the prototypes, which is defined by the latent feature satisfying the following conditions:

$$\mu_{c,k} = \text{argmin}_{z \in Z_c} ||z - o_{c,k}||_2^2, \tag{8}$$

where $o_{c,k}$ denotes the centroid of the k-th cluster for class c. The choice of K represents a trade-off between mode coverage and dictionary redundancy. While a larger K provides finer granularity, it increases the risk of cluster collapse or overfitting to noise. Note that the ablation study about K is shown in Table 3.

Algorithm 1 ProtoDiffAD: Training Procedure With MPG

Input: Normal training dataset $\mathcal{D}$, Pre-trained VAE Encoder $\mathcal{E}_\phi$, Number of prototypes K, Diffusion timesteps T, Unconditional probability p_{uncond}.
Output: Trained Denoising Network ϵ_θ,Prototype Dictionary $\mathcal{P} \in \{\mathcal{P}_c\}_{c=1:C_{\text{cls}}}$, where $\mathcal{P}_c$ indicates the set of prototypes for the c-class, and Class-wise projection network W_{Qf}^c
// Phase 1: Construct Prototype Dictionary
Initialize $\mathcal{P}_c \leftarrow \emptyset$ for all classes c
for $c = 1$ to C_{cls} **do**
    Collect latents $\mathbf{Z}_c = \{\mathcal{E}_\phi(\mathbf{x}) \mid (\mathbf{x}, y) \in \mathcal{D}, y = c\}$
    Compute prototypes $\{\boldsymbol{\mu}_{c,k}\}_{k=1}^{K} \leftarrow$ K-Means($\mathbf{Z}_c, K$)
    Append $\{\boldsymbol{\mu}_c\}$ to $\mathcal{P}_c$
end for
// Phase 2: Train Diffusion Model
Initialize ϵ_θ
while not converged **do**
    Sample batch $(\mathbf{x}_0, y) \sim \mathcal{D}$
    Sample $t \sim$ Uniform$(1, T)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Compute noisy state $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $z_0 = \mathcal{E}_\phi(x)$
    // Context Conditioning
    Retrieve prototypes $\mathcal{P}_y \subset \mathcal{P}$ corresponding to class y
    Compute context $\mathbf{c}_{proto}(\mathbf{z}_t, y)$ using Eqs. (8) and (9)
    if $t < T/3$ **then** w in Eqs. (8) and (9) is uniform distribution
    // CFG Dropout
    Sample $r \sim$ Uniform$(0, 1)$
    if $r < p_{\text{uncond}}$ **then** $\mathbf{c}_{proto} \leftarrow \emptyset$
    // Optimization
    Loss $\mathcal{L} = \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{proto})\|^2$
    Update θ via gradient descent $\nabla_\theta \mathcal{L}$
end while

2) MULTI-PROTOTYPE GUIDANCE (MPG)

Standard conditioning uses a static label y. However, forcing a generic class condition on a specific instance may lead to suboptimal reconstruction. We introduce an attention-based mechanism that computes a context-dependent mixture of prototypes.

a: PROTOTYPE CONTEXT AGGREGATION

Given a noisy latent z_t at timestep t and class y, we retrieve the corresponding prototypes $\mathcal{P}_y = \{\mu_{y,1}, \cdots, \mu_{y,K}\}$. To determine the relevance of each prototype to the current noisy state, we project z_t into a query vector using a learnable linear projection W_Q^y for the object class y. The relevance weights w_k are computed via softmax-normalised cosine similarity, defined as follows:

$$w_k = \frac{\exp(\text{sim}(W_Q^y z_t, \mu_{y,k})/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(W_Q^y z_t, \mu_{y,j})/\tau)}, \tag{9}$$

where τ is a temperature hyperparameter. The dynamic context vector c_{proto} is then obtained as the weighted sum:

$$c_{proto}(z_t, y) = \sum_{k=1}^{K} w_k \mu_{y,k}. \tag{10}$$

This attention mechanism ensures that the condition c_{proto} aligns with the normal mode most similar to the input, effectively guiding the denoising process towards the nearest valid normal distribution.

At high diffusion timesteps during the initial reverse process, the latent representation z_t is heavily saturated with Gaussian noise. In this low signal-to-noise regime, cosine similarity measurements become geometrically unstructured, rendering the dynamic context retrieval ineffective. To circumvent this instability, we assign a uniform probability distribution to the relevance weights w for the first $T/3$ steps of the denoising trajectory. Following this early stabilisation phase, the model transitions to dynamically computing the contextual relevance weights via Eq. (9).

b: PROTOTYPE GUIDANCE

We incorporate this dynamic context based on the principle of Classifier-Free Guidance [38]. During training, c_{proto} is randomly dropped with a probability of p_{uncond} to learn an unconditional score. During inference, the modified noise prediction is given by:

$$\tilde{\epsilon}_\theta(z_t, c_{proto}) = \epsilon_\theta(z_t, t, \emptyset) + s \cdot (\epsilon_\theta(z_t, t, c_{proto}) - \epsilon_\theta(z_t, t, \emptyset)), \tag{11}$$

where s is the guidance scale. By setting $s > 1.0$, we put more weight on reconstructing the samples to be more similar to the normal prototypes retrieved. Since anomalies fall outside this manifold represented by the prototypes, they incur larger reconstruction errors, facilitating localisation. Algorithm 1 describes the entire training procedure.

D. ANOMALY DETECTION

To compute the anomaly score α, we combine pixel-level and latent-level reconstruction discrepancies. Pixel-space residuals (*e.g.,* ℓ_1 or ℓ_2 distances between x_0 and $\tilde{x}_0$) are widely used because they yield spatially aligned anomaly maps that are easy to interpret. However, in highly textured regions, normal variability in fine-scale patterning and illumination can yield elevated pixel residuals even in defect-free areas (*e.g.,* repetitive or stochastic textures such as tile, carpet/grid patterns, wood grain, or woven textiles), thereby increasing false positives or reducing contrast for subtle defects. To address issues, latent-space discrepancies are presented. This approach can emphasise highly semantic deviations and may capture anomalies that induce weak pixel differences; however, because the latent representation is abstracted, very fine-scale defects may be attenuated.

Based on this principle, we employ a combination of the image-level and latent feature-level discrepancies presented by Beizaee et al. [6]. In their ablation study, the combination of two discrepancies achieves better distinguishing power than using them separately or simply combining them by arithmetic mean. This fusion strategy is particularly advantageous in robotic scenarios, where normal background textures (*e.g.,* wood grain or woven patterns) often produce elevated pixel-level residuals and may lead to false anomaly detections.

The abnormal score $\boldsymbol{a}$ is defined via the geometric mean of the clipped latent-level and pixel-level discrepancies:

$$\boldsymbol{a} = \sqrt{\frac{1}{\gamma_l \gamma_p} \min\left(\|\tilde{z}_0 - z_0\|, \gamma_l\right) \cdot \min\left(\|\tilde{x}_0 - x_0\|, \gamma_p\right)} \tag{12}$$

where γ_l and γ_p are truncation thresholds for the latent and pixel residuals, respectively, introduced to prevent excessively large deviations in either domain from dominating the fused score.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS AND EVALUATION PROTOCOLS

We conduct an extensive evaluation using two publicly available benchmark repositories for anomaly detection. Firstly, the MVTec-AD dataset [7] is employed. The MVTec-AD dataset comprises 5,354 high-fidelity images, partitioned into five textural categories and ten distinct object classes across varied domains. Secondly, we leverage the VisA dataset [8], which contains 10,821 high-resolution samples across 78 anomalous variations. The VisA dataset is structured into twelve subsets, categorised by the geometric and spatial complexity of the subjects, namely: single instances, multiple instances, and complex structural arrangements. It should be noted that all ablation studies are systematically performed on the MVTec-AD benchmark to ensure a consistent baseline for analysis. For fair training and testing, we use predefined splits used in other UAD studies [5], [6], [39], [40], [41] for training and testing.

2) EVALUATION METRICS

In alignment with established benchmarks in the field [6], [40], [41], the quantitative assessment of our model is predicated on a multi-faceted evaluation protocol spanning both global classification and local segmentation. We report the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) (*a.k.a.* Average Precision (AP)) to rigorously quantify performance at the image level (detection) and the pixel level (localisation). To account for the nuances of spatial accuracy in industrial defect identification, the Area Under the Per-Region-Overlap (AUPRO) is integrated as a primary metric for pixel-wise localisation. AUPRO measures the average overlap ratio between anomaly detection (AD) results and ground truth (GT) regions across the test split.

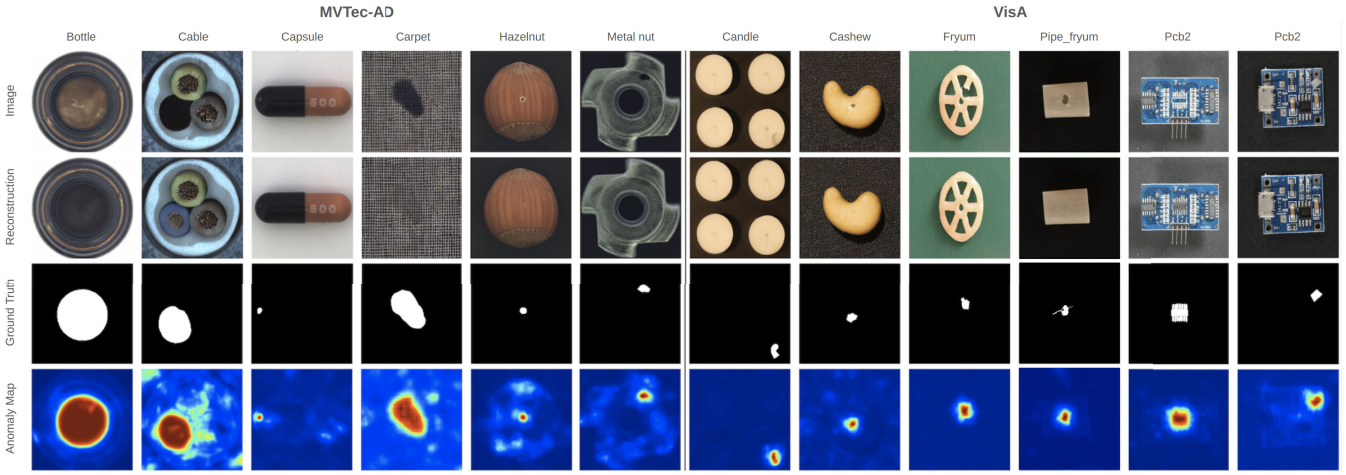


FIGURE 1. Qualitative results of ProtoDiffAD on the MVTec-AD and VisA datasets.

TABLE 1. Quantitative results on the VisA dataset.

Method	Image-level		Pixel-level		
	AUROC	AUPRC	AUROC	AUPRC	AUPRO
RD4AD '22	94.6	96.5	96.1	48.6	91.1
UniAD '22	96.5	98.8	96.8	43.4	90.7
SimpleNet '23	95.3	98.4	96.9	45.9	86.5
DeSTSeg '23	89.2	95.5	93.1	54.3	64.8
DiAD '24	97.2	99.0	96.8	52.6	90.7
MoEAD '24	98.0	99.3	96.9	49.8	91.0
GLAD '24	97.5	99.1	97.4	60.8	93.0
MambaAD '24	98.6	99.6	97.7	56.3	93.1
VPDM '24	98.4	-	97.8	-	-
ToCoAD '25	99.1	-	98.2	-	-
CRAS '25	98.3	-	98.0	-	-
DeCo-Diff '25	99.3	**99.8**	98.4	74.9	94.9
SGCAD '26	98.4	-	**98.7**	-	95.2
PatchCore '22	99.1	-	98.1	-	-
PaDiM '21	97.9	-	91.2	-	92.1
ProtoDiffAD	**99.5**	99.6	98.6	**78.1**	**95.3**

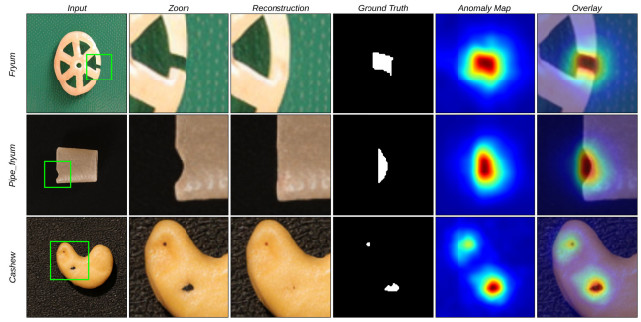


FIGURE 2. Anomaly detection results of ProtoDiffAD for samples having a tiny-sized anomaly area. The VisA dataset is used for the experiments.

3) IMPLEMENTATION DETAILS

ProtoDiffAD is trained with two NVIDIA RTX A6000 GPUs. We employ a pre-trained VAE of LDM [36]. The encoder $\mathcal{E}_\phi$ of the VAE is used to construct the prototype dictionary $\mathcal{P}$. Input images are centre-cropped and resized to 256 × 256, resulting in a compact latent resolution of $32\times32\times4$. The denoising backbone follows a standard U-Net architecture with channel multipliers of (1, 2, 4, 8) and spatial self-attention mechanisms at resolutions of 16×16 and 8×8.

The training procedure consists of two phases. First, we construct the prototype dictionary by clustering extracted latent features using K-Means++ [42] with a maximum of $K = 5$ centroids per class. K-means++ is employed for computational efficiency and stability. We select $K = 3$ as a robust default to sufficiently capture high intra-class variance without overfitting. Second, we optimise the diffusion model using the AdamW optimiser [43] ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate of 1×10^{-4} and cosine annealing schedule. A substantial global batch size of 64 is utilised to stabilise gradient estimation across the diverse unified multi-class distribution. To facilitate MPG, we set the guidance scale to $s = 2.0$, empirically determined to offer the optimal trade-off between anomaly suppression and structural fidelity (see Table 5).

TABLE 2. Quantitative results on the VisA dataset.

Method	Image-level		Pixel-level		
	AUROC	AUPRC	AUROC	AUPRC	AUPRO
RD4AD '22	92.4	92.4	98.1	38.0	91.8
UniAD '22	88.8	90.8	98.3	33.7	85.5
SimpleNet '23	87.2	87.0	96.8	34.7	81.4
DeSTSeg '23	88.9	89.0	96.1	39.6	67.4
DiAD '24	86.8	88.3	96.0	26.1	75.2
MoEAD '24	93.0	95.1	98.7	36.6	88.6
GLAD '24	91.8	92.9	97.4	34.6	92.0
MambaAD '24	94.3	94.5	98.5	39.4	91.0
VPDM '24	94.2	-	**98.9**	-	-
ToCoAD '25	95.4	-	98.4	-	-
CRAS '25	93.5	95.2	97.7	38.4	-
DeCo-Diff '25	96.4	96.8	98.5	51.3	92.1
SGCAD '26	93.4	-	98.8	-	89.8
ProtoDiffAD	**97.1**	**97.5**	98.6	**55.6**	**93.8**

TABLE 3. ProtoDiffAD results with respect to the conditioning term and prototype dictionary size (K). Results for prototype-guided diffusion are reported as mean ± standard deviation over five independent K-Means++ initialisations. Note that Uncond. means 'unconditioned' and Cls. embed. means 'class embedding'.

Condition	Image-level		Pixel-level		
	AUROC	AUPRC	AUROC	AUPRC	AUPRO
Uncond.	94.1	96.4	96.2	51.6	83.2
Cls. embed.	95.8	96.5	93.7	52.4	91.2
With prototype-guided diffusion (K scale)					
$K = 1$	97.3 ± 0.00	98.7 ± 0.00	96.8 ± 0.00	75.3 ± 0.00	94.2 ± 0.00
$K = 2$	99.3 ± 0.06	99.4 ± 0.04	98.1 ± 0.07	76.8 ± 0.18	94.7 ± 0.10
$K = 3$	$\mathbf{99.5 \pm 0.04}$	$\mathbf{99.6 \pm 0.03}$	$\mathbf{98.6 \pm 0.05}$	$\mathbf{78.1 \pm 0.14}$	95.3 ± 0.08
$K = 5$	99.4 ± 0.05	$\mathbf{99.6 \pm 0.04}$	$\mathbf{98.6 \pm 0.06}$	77.4 ± 0.17	$\mathbf{95.4 \pm 0.09}$
$K = 10$	99.4 ± 0.08	99.5 ± 0.06	98.2 ± 0.09	76.8 ± 0.23	95.1 ± 0.13

Additionally, $\tau = 0.5$ is applied, and p_{uncond} is set by 0.2. The training epoch is set to 1000 for each dataset. In inference, we utilise the Denoising Diffusion Implicit Models (DDIM) [44] sampler with $T = 20$ steps to accelerate generation. γ_p and γ_l are set to 0.4 and 0.2, respectivley. These values are taken from our ablation studies and the suggestion from Beizaee et al. [6], which present the anomaly scoring metrics.

B. COMPARISON WITH EXISTING METHODS

We compare our approach to nine recently-proposed approaches for multi-class UAD: RD4AD [3], UniAD [5], SimpleNet [45], DeSTSeg [2], DiAD [40], MoEAD [46], GLAD [29], MambaAD [41], VPDM [34], ToCoAD [19], CRAS [20], DeCo-Diff [6], SGCAD [25]. PatchCore [4], and PaDiM [16]. Our comparison covers performance not only between the multi-class UDA methods, but also a comparison with the 'one-class-one-model' approaches [4], [16]

1) QUANTITATIVE RESULTS

The results for the MVTec-AD dataset are detailed in Table 1. Despite the performance saturation often observed on this benchmark, ProtoDiffAD establishes a new state-of-the-art. We report an Image-AUROC of 99.5%, outperforming the nearest competitor. In terms of pixel-level segmentation, our method demonstrates superior fidelity, achieving a competitive Pixel-AUROC score (98.6%) and the highest Pixel-AUPRC (78.1%) and Pixel-AUPRO (95.3%) scores. Notably, the Pixel-AUPRC score represents a relative improvement of roughly 4.3% over the second-best method (74.9%). This suggests that ProtoDiffAD effectively minimises false positives while maintaining high sensitivity to defect regions. An interesting result is the comparison with the methods taking a *one-class-one-model* approach. PatchCore [4] and PaDiM [16] are compiled AD models for each class. PatchCore and PaDiM achieve image-level AUROCs of 99.1% and 97.9%, which are lower than that of ProtoDiffAD. Qualitative results on the MVTec-AD dataset show that ProtoDiffAD can outperform the class-wise AD methods.

Table 2 presents the results on the VisA dataset, which presents a challenging environment due to its complex object structures and multi-class diversity. ProtoDiffAD achieves a leading Image-AUROC of 97.1% and an Image-AUPRC of 97.5%. A marked advantage is observed in the pixel-level localisation tasks, particularly within the precision-recall trade-off. Our method attains a Pixel-AUROC of 98.6%, close to the best performer, VPDM, which achieves 98.9%. This indicates that prototypical guidance significantly aids in delineating anomalies from complex backgrounds.

Consequently, the quantitative evidence indicates that ProtoDiffAD provides robust generalisation capabilities. By leveraging the generative priors of diffusion models alongside prototypical representations, the method delivers precise anomaly localisation, as evidenced by substantial gains in AUPRC and AUPRO across both datasets.

2) QUALITATIVE RESULTS

Fig. 1 presents anomalous samples from diverse categories in the MVTec-AD and VisA datasets, along with their corresponding reconstructions, ground-truth masks, and predicted anomaly maps. As illustrated in the second row (Reconstruction), our prototype-guided diffusion model demonstrates a robust capability to restore anomalous inputs to a nominal state whilst preserving the high-frequency details of the non-defective background. This behaviour validates the effectiveness of the MPG. By conditioning the reverse diffusion process on a context-dependent mixture of normal prototypes, the generation trajectory is constrained to reside within the learnt manifold of normal samples. Consequently, anomalous features, which lack corresponding representations in the prototype dictionary, are effectively corrected during reconstruction.

For instance, in the Hazelnut and Cable classes (MVTec-AD), the model successfully removes surface cracks and structural deformations, replacing them with coherent normal textures. Similarly, in the complex PCB and Fryum scenarios (VisA), the model accurately hallucinates missing components or corrects geometric irregularities to match the standard configuration. The resulting anomaly maps (fourth row) exhibit high contrast, with activation regions aligning precisely with the ground truth masks (third row). These circumstances are also observed from zoomed results represented in Fig. 2. ProtoDiffAD successfully reconstructs small anomaly areas, so it can compute a large reconstruction error. This indicates that ProtoDiffAD effectively suppresses false positives in complex backgrounds whilst maintaining the sensitivity required for precise anomaly localisation in a unified multi-class setting.

C. ABLATION STUDIES

1) PROTOTYPE DICTIONARY

We construct a per-class prototype dictionary by clustering normal latent features, where the dictionary scale K specifies the number of centroids retained for each class. Table 3 reports the impact of the conditioning term and varying K on MVTec-AD. For the prototype-guided configurations,

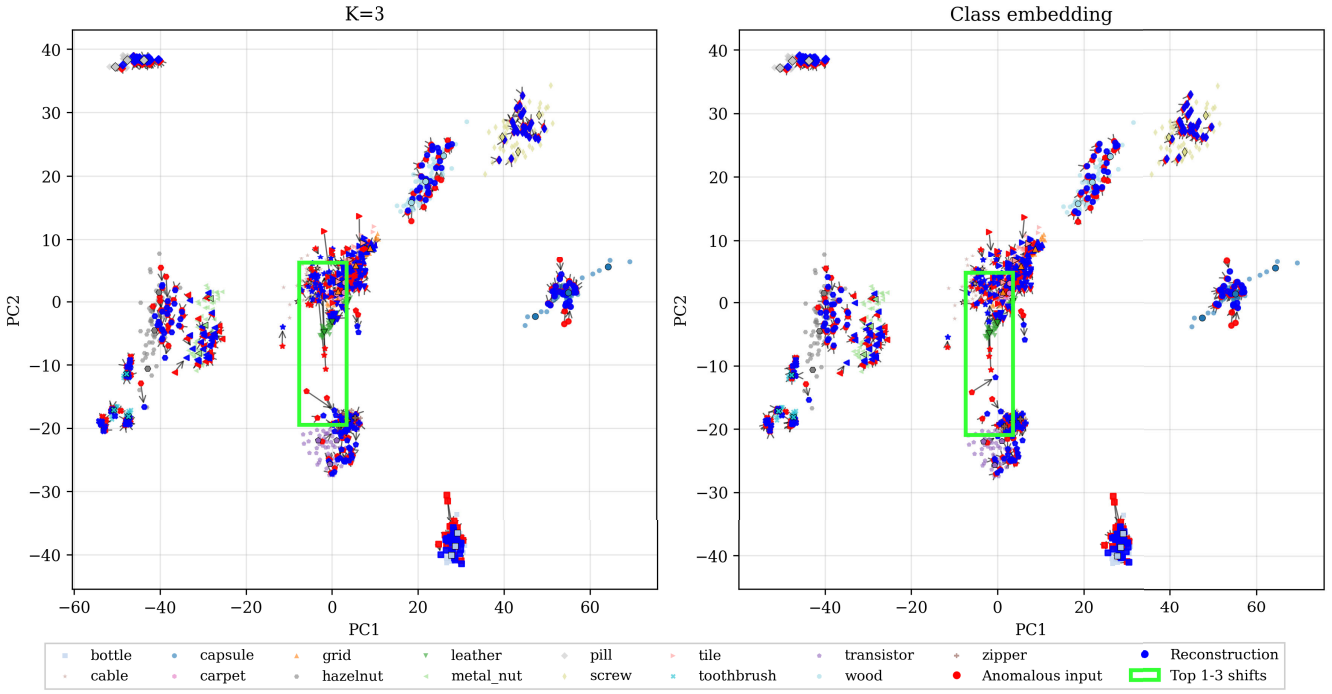


FIGURE 3. Latent feature distribution of each category depending on the conditioning approach on the MVTec-AD datasets. We compare 'Class embedding', which is a naive LDM conditioned with explicit class context, and ProtoDiffAD with $K = 3$. We apply PCA to the output of $\mathcal{E}_\phi(x)$ for dimension reduction. The 15 samples that produce the top 15 differences between anomalous and reconstruction results are connected by arrow lines.

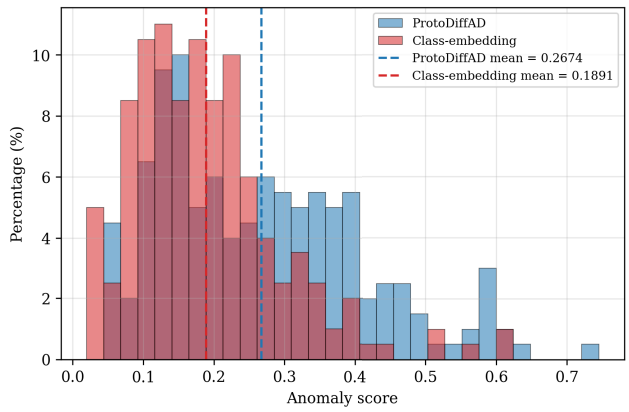


FIGURE 4. Anomaly score distribution of the 'Class embedding' with LDM and ProtoDiffAD.

the results are reported as mean ± standard deviation over three independent K-Means++ initialisations, while the remaining training and inference settings are held fixed. Without conditioning, the model attains an Image-AUROC of 94.1% and a Pixel-AUPRC of 51.6%. Replacing the prototype conditioning with a standard class embedding increases Image-AUROC to 95.8%; however, Pixel-AUROC decreases to 93.7%, suggesting that coarse class conditioning does not consistently benefit localisation in this setting.

Introducing prototype-guided conditioning improves performance across all reported metrics for the examined dictionary sizes. With a single prototype per class ($K = 1$), Image-AUROC increases to 97.3% and Pixel-AUPRC to 75.3%. Increasing the dictionary size further yields additional gains: $K = 2$ achieves 99.3% Image-AUROC and 76.8% Pixel-AUPRC, and $K = 3$ achieves 99.5% Image-AUROC, 78.1% Pixel-AUPRC, and 95.3% Pixel-AUPRO. Moving to $K = 5$ results in similar performance (*i.e.,* 99.4% Image-AUROC and 95.4% Pixel-AUPRO), indicating that, within the explored range, improvements tend to saturate beyond a moderate number of prototypes. The standard deviations remain small across the examined configurations. For $K = 3$, the standard deviations are 0.04, 0.03, 0.05, 0.14, and 0.08 percentage points for Image-AUROC, Image-AUPRC, Pixel-AUROC, Pixel-AUPRC, and Pixel-AUPRO, respectively. The $K = 2$ and $K = 5$ configurations show similarly limited variation. For $K = 1$, no measurable variation is observed because the single-cluster solution yields the same class centroid for a fixed latent feature set. At $K = 10$, the variability increases slightly, with the largest standard deviation reaching 0.23 percentage points for Pixel-AUPRC, which may reflect the greater number of possible partitions when more prototypes are retained. Overall, these results suggest limited sensitivity to K-Means++ initialisation within the explored range.

The latent feature distribution illustrated in Fig. 3 and the distribution of anomaly scores in Fig. 4 also suggest that conditioning by prototypes shows better reconstruction results compared with explicit class embedding. The black arrow lines in Fig. 3 show the positional change of the anomalous sample after reconstruction. As shown in the

TABLE 4. Quantitative results of ProtoDiffAD depending on the CFG dropout probability p_{uncond} on MVTec-AD.

p_{uncond}	Image-level		Pixel-level		
	AUROC	AUPRC	AUROC	AUPRC	AUPRO
0.00	97.9	98.7	96.8	66.8	92.1
0.05	98.8	99.2	97.7	73.6	94.0
0.10	99.3	99.5	98.6	78.0	94.9
0.20	**99.5**	**99.6**	**98.6**	**78.1**	**95.3**
0.30	99.4	99.6	98.5	77.9	95.1
0.50	98.9	99.2	97.9	74.2	94.1

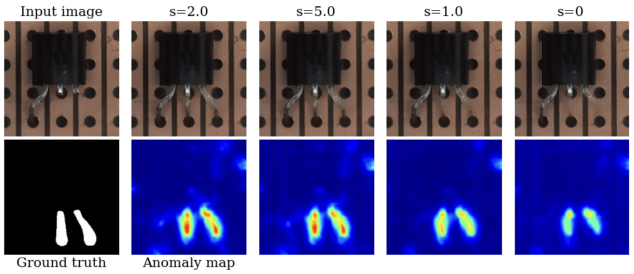


FIGURE 5. Qualitative results of anomaly detections of ProtoDiffAD depending on the guidance scale *s*.

TABLE 5. ProtoDiffAD results with respect to the guidance scale *s*.

Scale (s)	Image-level		Pixel-level		
	AUROC	AUPRC	AUROC	AUPRC	AUPRO
0	93.5	94.2	95.1	40.5	90.5
1.0	99.5	99.3	97.9	69.4	94.1
2.0	**99.5**	**99.6**	**98.6**	**78.1**	**95.3**
5.0	99.4	99.5	98.5	77.5	95.1

figure, the results of ProtoDiffAD show a more dramatic change, *i.e.,* after reconstructing anomalous samples, they are closer to the normal sample distribution (see the arrow lines in the green boxes). Additionally, as shown in Fig. 4, ProtoDiffAD provides discriminative power compared with usage of naive class embedding. Average anomaly score of ProtoDiffAD is 0.2674, and it is 41.4% higher than the 'Class embedding'. As a result, the reconstruction results of the ProtoDiffAD trained with $K = 3$ are closer to the normal sample distribution. We have set $K = 3$ for further experiments.

2) INFLUENCE OF CFG DROPOUT PROBABILITY P_{uncond}

Classifier-free guidance requires the denoising network to learn both prototype-conditioned and unconditional noise predictions. The dropout probability p_{uncond} controls the relative frequency of these training cases and may consequently affect the reliability of the guided prediction. We therefore evaluate $p_{\text{uncond}} \in \{0, 0.05, 0.10, 0.20, 0.30, 0.50\}$.

As reported in Table 4, setting $p_{\text{uncond}} = 0$ produces a substantial reduction in localisation performance. In this configuration, the unconditional prediction used by CFG is not explicitly optimised, reducing Pixel-AUPRC to 66.8%. Introducing unconditional dropout consistently improves performance, with a broad stable region between $p_{\text{uncond}} = 0.10$ and 0.30. The default $p_{\text{uncond}} = 0.20$ achieves the strongest overall performance, including 99.5% Image-AUROC, 78.1% Pixel-AUPRC, and 95.3% AUPRO.

Excessive dropout is also detrimental. At $p_{\text{uncond}} = 0.50$, Pixel-AUPRC decreases to 74.2%, suggesting that omitting the prototype context from half of the training cases weakens the model's ability to learn prototype-guided normality correction. We therefore retain $p_{\text{uncond}} = 0.20$, which provides the best balance among the evaluated settings and is located within a relatively stable operating range.

TABLE 6. Detection and localisation performance across varying softmax temperature (τ) scales.

Temp. (τ)	Image-level	Pixel-level		
	AUROC	AUROC	AUPRC	AUPRO
0.1	98.2	96.5	65.4	88.7
0.2	98.9	97.4	72.3	92.1
0.5	**99.5**	**98.6**	**78.1**	**95.3**
1.0	99.1	97.8	74.2	93.4
5.0	97.4	95.2	61.3	85.2

3) INFLUENCE OF GUIDANCE SCALE *s*

The impact of the guidance scale s of MPG on anomaly detection performance is detailed in Table 5. When $s = 0$ (unconditional generation), the model achieves a baseline Image-AUROC of 93.5%. However, this setting lacks the explicit constraints needed to reject anomalous patterns, resulting in a significantly lower Pixel-AUPRC of 40.5%. Increasing the scale to $s = 1.0$ (standard conditional) yields a substantial improvement, boosting Image-AUROC by 6.4% and Pixel-AUPRC by nearly 71% relative to the baseline. The performance peaks at $s = 2.0$, where Image-AUROC reaches 99.5%, and Pixel-AUPRC maximises at 78.1%.

Those results confirm that a moderate guidance strength effectively forces the reconstruction toward the learned normal prototypes, thereby amplifying the error in anomalous regions. Pushing the scale further to $s = 5.0$ yields diminishing returns, suggesting that while strong guidance is beneficial, excessive forcing may introduce high-frequency artefacts that slightly degrade structural consistency. As illustrated in Fig. 5, performance saturates beyond $s = 2.0$, indicating that a scale greater than 2.0 is necessary to effectively cover the normal manifold without over-segmenting the latent space.

4) TEMPERATURE SCALING IN MULTI-PROTOTYPE GUIDANCE

The parameter τ fundamentally controls the sharpness of the generative guidance [47]. Table 6 presents the detection and localisation performance across varying temperature scales.

TABLE 7. Inference latency and localisation performance across varying DDIM inference steps (T). I-/P- means Image- and Pixel-level, respectively.

DDIM Steps (T)	FPS	I-AUROC	P-AUROC	P-AUPRC
5	45	96.1	94.2	62.1
10	24	98.5	97.1	74.5
20	12	99.5	98.6	78.1
50	5	99.6	98.7	78.4

Setting $\tau = 0.1$ induces a near-hard assignment, allocating the majority of the guidance weight to a single prototype centroid. Whilst this enforces strict adherence to a specific manifold point, it simultaneously discards the nominal intra-class variance inherent to the input image, resulting in elevated false-positive rates in defect-free regions and a corresponding degradation in the Pixel AUPRC (65.4%). Conversely, as τ increases towards 5.0, the attention distribution approaches uniformity. This over-smoothing forces the dynamic context vector into a generic arithmetic mean, reintroducing the mode-averaging phenomenon and blurring fine-grained structural details. The experimental optimum at $\tau = 0.5$ strikes the necessary balance, permitting the diffusion model to interpolate smoothly between prototypes to preserve healthy variance whilst maintaining sufficient geometric constraint to expose anomalous deviations.

5) INFERENCE DYNAMICS AND COMPUTATIONAL PROFILING

To quantify the operational viability of the proposed framework, latency and localisation performance are evaluated as a function of the number of Denoising Diffusion Implicit Model (DDIM) inference steps (T). The profiling, conducted on a single NVIDIA RTX A6000 GPU, is detailed in Table 7.

Iterative denoising represents the primary computational bottleneck for diffusion architectures in industrial deployment [44]. The profiling indicates that extreme acceleration ($T = 5$) retains sufficient global semantic structure to maintain a respectable Image AUROC (96.1%), but fails to resolve the high-frequency residual details required for precise pixel-level segmentation. The default setting of 20 DDIM steps provides an operationally viable throughput of 12 Frames Per Second (FPS) whilst fully saturating the model's localisation capabilities. Extending the trajectory to 50 steps yields negligible performance gains, confirming that 20 steps under Multi-Prototype Guidance provides optimal convergence to the normal manifold without incurring unnecessary computational overhead.

6) EARLY STAGE WITH UNIFORM ATTENTION WEIGHTING

Table 8 evaluates the effectiveness of uniform weighting for early-stage denoising. When dynamic weighting is activated from the first reverse step (*a.k.a.* no uniform weighting), the model achieves an Image-AUROC of 99.1% but a lower Pixel-AUPRC of 74.8%. This result indicates that prototype similarities computed from highly noisy latent states are insufficiently reliable for precise localisation.

TABLE 8. Sensitivity to the duration of uniform prototype weighting. *'Early stage'* denotes the fraction of the reverse DDIM trajectory using uniform weights, and N_{uni} is the corresponding number of uniform-weighting steps for $N_{\text{DDIM}} = 20$.

Early stage	N_{uni} step	Image-level		Pixel-level		
		AUROC	AUPRC	AUROC	AUPRC	AUPRO
0	0	99.1	99.4	98.0	74.8	93.9
$T/4$	5	99.4	99.5	98.4	77.2	94.8
$\boldsymbol{T/3}$	**7**	**99.5**	**99.6**	**98.6**	**78.1**	**95.3**
$T/2$	10	99.3	99.5	98.3	76.4	94.6
T	20	98.7	99.0	97.4	70.6	92.7

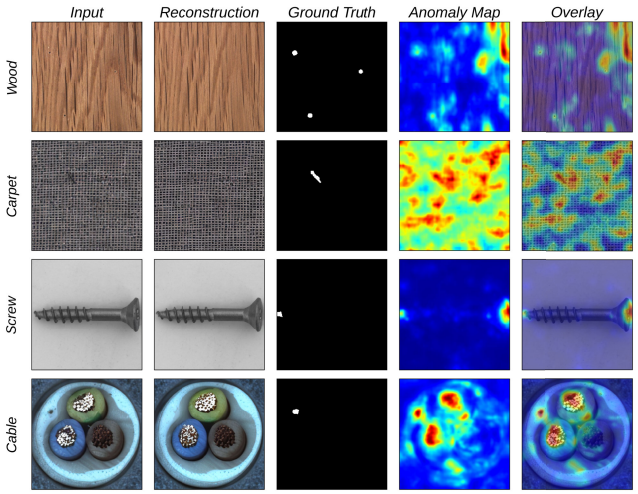


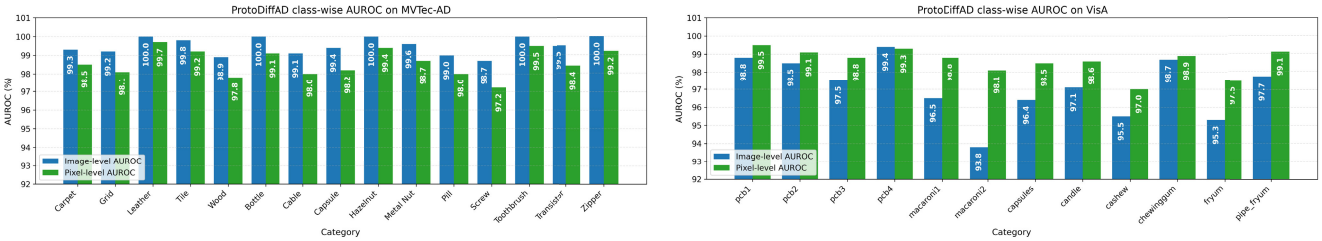
FIGURE 6. Examples of failure cases for ProtoDiffAD's anomaly detection.

Introducing a short uniform-weighting phase improves all localisation metrics. The $T/4$ configuration achieves a Pixel-AUPRC of 77.2% and an AUPRO of 94.8%, while the default $T/3$ configuration produces the strongest overall performance, reaching 78.1% of Pixel-AUPRC and 95.3% of AUPRO. Extending the uniform phase to $T/2$ causes a moderate decrease to 76.4% Pixel-AUPRC. This reduction suggests that delaying dynamic attention for too long prevents the model from exploiting informative instance-dependent prototype similarities during the intermediate denoising stages.

Uniform weighting throughout the complete trajectory yields the largest degradation, particularly in Pixel-AUPRC, because the prototype context becomes a fixed arithmetic average and loses its ability to select normal modes relevant to individual inputs. Overall, the results demonstrate a broad stable region between $T/4$ and $T/2$, indicating that ProtoDiffAD is not critically dependent on the exact boundary. We retain $T/3$ because it provides the strongest balance between early-stage stability and later-stage adaptive guidance.

D. LIMITATION

ProtoDiffAD remains sensitive to reconstruction uncertainty in highly textured regions. As shown in Fig. 6, stochastic



(a) MVTec-AD dataset.

(b) VisA dataset.

FIGURE 7. Class-wise image-level detection and pixel-level AUROCs of ProtoDiffAD on the MVTec-AD and VisA datasets.

or repetitive patterns in categories such as *carpet* and *wood* may not be reconstructed at precisely the same spatial phase, causing residual activation in defect-free regions. Conversely, very small or low-contrast defects may be attenuated by the compressed latent representation, leading to incomplete or spatially diffuse localisation near object boundaries. As shown in Fig. 7, the class-wise results are consistent with these limitations: on MVTec-AD, *wood* and *screw* show relatively low pixel-level AUROCs of 97.8% and 97.2%, while *carpet* achieves 98.5%. On VisA, *macaroni2*, *fryum*, and *cashew* exhibit lower image-level AUROCs despite retaining comparatively high pixel-level scores, suggesting that local anomaly evidence may remain detectable but insufficiently concentrated for reliable image-level decisions. Thus, combining latent- and pixel-level discrepancies mitigates, but does not eliminate, the trade-off between texture-induced false positives and sensitivity to fine defects.

Additionally, computational efficiency is also one of the main limitations of ProtoDiffAD. Even though the proposed ProtoDiffAD provides seamless AD performance, it is still computationally heavy and requires high-performance computational processes. The current inference profiling was conducted on a single NVIDIA RTX A6000 GPU. At $T = 20$, ProtoDiffAD achieves 12 FPS, corresponding to approximately 83 ms per image. This throughput may support moderate-rate or asynchronously triggered inspection, but it should not be interpreted as universally satisfying real-time requirements for high-speed industrial production lines. Moreover, performance on consumer-grade GPUs and edge devices has not yet been evaluated. Future work will investigate accelerated sampling, diffusion-step reduction, model distillation, quantisation, pruning, and hardware-specific optimisation for resource-constrained deployment.

V. CONCLUSION AND DISCUSSION

We have presented ProtoDiffAD, a novel prototype-guided diffusion framework tailored for unified multi-class unsupervised anomaly detection. We have introduced a reformulated Multi-Prototype Guidance (MPG), which dynamically aggregates relevant prototypes to guide the diffusion process. By forcing the generative model to strictly adhere to the nearest normal modes during reconstruction, ProtoDiffAD effectively amplifies the reconstruction error in anomalous regions while preserving high-fidelity details in normal areas. Our comprehensive experiments have demonstrated that ProtoDiffAD successfully captures complex intra-class variances and achieves state-of-the-art performance compared to existing benchmarks in the unified multi-class setting.

Whilst ProtoDiffAD demonstrates competitive, and frequently superior, performance relative to contemporary state-of-the-art baselines, it possesses a notable theoretical limitation. The framework constructs its prototype dictionary by applying the standard K-Means clustering algorithm to latent representations extracted by the VAE's encoder. K-Means relies on minimising Euclidean distances, which inherently fails to accurately characterise the intrinsic geometric structure of complex manifolds. Furthermore, the algorithm operates under the restrictive assumption that the data distribution is spherical. In complex real-world scenarios, this assumption is fundamentally unrealistic. Enforcing such a constraint frequently yields locally optimised centroids that fail to encapsulate the true underlying data manifold, consequently degrading the overall anomaly detection (AD) performance. In our future work, we aim to address this by developing novel prototype formulations that explicitly accommodate the non-linear and highly intricate distributions of real-world data, thereby further advancing AD capabilities.

REFERENCES

[1] H. Yao, Y. Cao, W. Luo, W. Zhang, W. Yu, and W. Shen, "Prior normality prompt transformer for multiclass industrial image anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 10, pp. 11866–11876, Oct. 2024.

[2] X. Zhang, S. Li, X. Li, P. Huang, J. Shan, and T. Chen, "DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3914–3923.

[3] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9737–9746.

[4] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 14318–14328.

[5] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst. 35*, 2022, pp. 4571–4584.

[6] F. Beizaee, G. A. Lodygensky, C. Desrosiers, and J. Dolz, "Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 19088–19097.

[7] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.

[8] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 392–408.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[10] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8642–8651.

[11] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.

[12] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 146–157.

[13] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 3110–3118.

[14] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, "Abnormal event detection and localization via adversarial event prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3572–3586, Aug. 2022.

[15] J. Yu, J.-G. Kim, J. Gwak, B.-G. Lee, and M. Jeon, "Abnormal event detection using adversarial predictive coding for motion and appearance," *Inf. Sci.*, vol. 586, pp. 59–73, Mar. 2022.

[16] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PADIM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 475–489.

[17] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9664–9674.

[18] V. Zavrtanik, M. Kristan, and D. Skočaj, "DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Apr. 2021, pp. 8330–8339.

[19] Y. Liang, Z. Hu, J. Huang, D. Di, A. Su, and L. Fan, "ToCoAD: Two-stage contrastive learning for industrial anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[20] Q. Chen, H. Luo, H. Yao, W. Luo, Z. Qu, C. Lv, and Z. Zhang, "Center-aware residual anomaly synthesis for multiclass industrial anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 21, no. 9, pp. 7276–7286, Sep. 2025.

[21] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8032–8041.

[22] E. Iqbal, S. U. Khan, S. Javed, B. Moyo, Y. Zweiri, and Y. Abdulrahman, "Multi-scale feature reconstruction network for industrial anomaly detection," *Knowl.-Based Syst.*, vol. 305, Dec. 2024, Art. no. 112650.

[23] M. Lee and J. Choi, "Text-guided variational image generation for industrial anomaly detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26519–26528.

[24] Y. Qiang, J. Cao, S. Zhou, J. Yang, L. Yu, and B. Liu, "TGARD: Text-guided adversarial reconstruction for industrial anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 21, no. 12, pp. 9297–9308, Dec. 2025.

[25] Y. Zhong, M. Lei, X. Zhong, B. Guo, and S. Dian, "Semantically guided counterfactual model for multiclass anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 22, no. 6, pp. 1–12, Jun. 2026.

[26] R. Zhang, H. Wang, M. Feng, Y. Liu, and G. Yang, "Dual-constraint autoencoder and adaptive weighted similarity spatial attention for unsupervised anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 7, pp. 9393–9403, Jul. 2024.

[27] W. Rahmaniar and K. Suzuki, "Multi-AD: Cross-domain unsupervised anomaly detection for medical and industrial applications," *Pattern Recognit.*, vol. 172, Apr. 2026, Art. no. 112486.

[28] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, and L. Xie, "DiAD: A diffusion-based framework for multi-class anomaly detection," 2023, *arXiv:2312.06607*.

[29] H. Yao, M. Liu, H. Wang, Z. Yin, Z. Yan, X. Hong, and W. Zuo, "Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024.

[30] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[31] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[32] J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi, "Prototypical contrastive learning of unsupervised representations," 2020, *arXiv:2005.04966*.

[33] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[34] Y. Li, Y. Feng, B. Chen, W. Chen, Y. Wang, X. Hu, B. Sun, C. Qu, and M. Zhou, "Vague prototype-oriented diffusion model for multi-class anomaly detection," in *Proc. 41st Int. Conf. Mach. Learn.*, vol. 235, Jul. 2024, pp. 27771–27790.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.

[37] J. Yu, H. Oh, Y. Lee, and J. Yang, "Denoising diffusion model with adversarial learning for unsupervised anomaly detection on brain MRI images," *Pattern Recognit. Lett.*, vol. 186, pp. 229–235, Oct. 2024.

[38] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022, *arXiv:2207.12598*.

[39] J. Yu, H. Oh, M. Kim, and J. Kim, "Normality-calibrated autoencoder for unsupervised anomaly detection on data contamination," 2021, *arXiv:2110.14825*.

[40] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, and L. Xie, "A diffusion-based framework for multi-class anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 8472–8480.

[41] H. He, Y. Bai, J. Zhang, Q. He, H. Chen, Z. Gan, C. Wang, X. Li, G. Tian, and L. Xie, "MambaAD: Exploring state space models for multi-class unsupervised anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 71162–71187.

[42] D. Arthur et al., "K-means++: The advantages of careful seeding," *Soda*, vol. 7, pp. 1027–1035, 2007.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[44] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[45] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A simple network for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20402–20411.

[46] S. Meng, W. Meng, Q. Zhou, S. Li, W. Hou, and S. He, "MoEAD: A parameter-efficient model for multi-class anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.



JONGMIN YU received the first Ph.D. degree from the School of Electrical Engineering and Computer Science, GIST, Gwangju, South Korea, and the second Ph.D. degree from the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, WA, Australia. He was a Research Associate with DAMTP, University of Cambridge. He was with King's College London and the Institute of IT Convergence, Korea Advanced Institute of Science and Technology. He is currently a Principal Scientist with nextstudio Inc. His research interests include artificial intelligence, machine learning, pattern recognition, and the mathematical understanding of these.



HYEONTAEK OH received the B.S. degree in computer science (summa cum laude) and the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), in 2012 and 2020, respectively. From 2020 to 2023, he was a Team Leader with the KAIST Institute for Information Technology Convergence. He has been the Head of the AI Laboratory, nextstudio Inc., since 2024. His research interests include the applications of artificial intelligence and agent systems.



ZHONGTIAN SUN was a Research Associate with the University of Cambridge and the University of Oxford. He is currently a Lecturer (an Assistant Professor) with the School of Computing, University of Kent, and a Visiting Fellow with the Department of Computer Science and Technology, University of Cambridge. He also has industry experience in asset management and is the Co-Founder and CTO of an AI4finance startup. His research focuses on enhancing the expressiveness and interpretability of deep learning models, with applications including healthcare, finance, education, and recommender systems. His broader interests include representation learning, large language models, multi-agent systems, causal inference, and symbolic reasoning.



ANOUSHKA HARIT received the bachelor's and master's degrees in computer science. She was a Research Software Engineer with DeepMind, where she contributed to large-scale machine learning systems. She is currently a Ph.D. Researcher with the University of Cambridge, focusing on the intersection of artificial intelligence, machine learning, and healthcare. Her research focuses on developing reliable and interpretable AI systems for medical applications, with particular emphasis on cancer detection and clinical decision support. Her current interests include medical AI, AI safety, representation learning, and trustworthy machine learning.



NAMHYUN YOO (Member, IEEE) received the M.S. and Ph.D. degrees in intelligent agent and semantic web from Sunchon National University, South Korea, in 1999 and 2007, respectively. From 2008 to 2010, he conducted research on cooperative intelligence with the Artificial Intelligence Super Laboratory, The University of Oklahoma, USA. Since 2011, he has been with Kyungnam University, Changwon, South Korea, where he is currently a Professor with the School of Computer Science and Engineering. His research interests include context awareness, digital twin, and software engineering for smart manufacturing.



JINHONG YANG (Member, IEEE) received the Ph.D. degree in information and communication engineering from the Information and Communication Engineering Department, KAIST, Daejeon, in 2017. From 2005 to 2008, he was a Research Assistant with Herit Inc. He is currently an Associate Professor with the Department of Medical IT, Inje University. He holds more than 300 domestic registered patents and over 30 international registered patents in the field of information and communications technology. His research interests include reinforcement learning, physical AI, and embodied AI.