Model Selection of Anomaly Detectors in the Absence of Labeled Validation Data

Clement Fung, Chen Qiu, Aodong Li, and Maja Rudolph

***Abstract*—Anomaly detection is the task of identifying abnormal samples in large unlabeled datasets. Although the advent of foundation models has produced powerful zero-shot anomaly detection methods, their deployment in practice is often hindered by the absence of labeled validation data—without it, detection performance cannot be evaluated reliably. In this work, we propose SWSA (Selection With Synthetic Anomalies): a general-purpose framework to select image-based anomaly detectors without labeled validation data. Instead of collecting labeled validation data, we generate synthetic anomalies from a small support set of normal images without using any training or fine-tuning. Our synthetic anomalies are then used to create detection tasks that compose a validation framework for model selection. In an empirical study, we evaluate SWSA with three types of synthetic anomalies and on two selection tasks: model selection of image-based anomaly detectors and prompt selection for CLIP-based anomaly detection. SWSA often selects models and prompts that match selections made with a ground-truth validation set, outperforming baseline selection strategies.**

***Impact Statement*—Foundation models can be applied to a variety of anomaly detection tasks with little or no additional training data. However, a lack of labeled anomalies still hinders the deployment of these models, as their detection accuracy cannot be validated. In this work, we propose a method for selecting anomaly detectors that only requires a small amount of benign data. We use our method to select image-based anomaly detection models and show that it often selects the true best-performing models and configurations. Our method reduces the labor-intensive and economic burden of collecting representative datasets of anomalies for validation, enabling model selection with reduced data requirements.**

***Index Terms*—Artificial intelligence algorithmic design and analysis, Testing machine learning, Unsupervised learning**

I. Introduction

Anomaly detection, identifying samples that deviate from normal behavior, is an important task for supporting medical diagnosis [19], financial transactions [1], cybersecurity [47], [66], and industrial operations [4]. Recent developments in foundation models suggest that it is possible to pre-train a model on a large dataset from one domain and deploy it for new anomaly detection tasks [41], [42], [80], providing the exciting possibility to deploy anomaly detectors for new applications without training data. However, one must trust that foundation models perform as expected before deployment. Validating the performance of anomaly detection models is often hindered by the absence of labeled validation data, since anomalies are, by definition, rare [22], [69].

In this work, we propose to use synthetic anomalies for selecting image-based anomaly detectors through our proposed framework: SWSA (Selection With Synthetic Anomalies). We compare two promising types of strategies for generating synthetic anomalies: (i) data augmentation methods [39] and (ii) style transfer with pre-trained diffusion models [29]. Our methods assume access to only a small support set of normal images and do not require any training, fine-tuning, or domain-specific techniques. We then label these generated images as anomalies to create synthetic validation datasets for selecting candidate anomaly detection models with SWSA. We find that SWSA often matches the selections made with real validation sets and outperforms baseline selection strategies across a variety of anomaly detection tasks and domains, ranging from natural images to industrial defects.

Our work makes the following contributions:

- In Sec. III-A, we propose SWSA: a framework to select anomaly detection models with synthetic anomalies. Fig. 1 shows the outline of our approach.
- In Sec. III-B, we propose a practical technique for generating synthetic anomalies with a general-purpose pre-trained diffusion model—without any fine-tuning or auxiliary datasets. When used in SWSA, we show that these synthetic anomalies are most effective for model selection in natural settings (i.e., birds and flowers).
- In Sec. IV, we empirically evaluate SWSA with a variety of anomaly-generation methods, datasets, and anomaly-detection tasks. We show that SWSA is effective in two use cases: model selection from a set of candidate anomaly detectors (Sec. IV-B) and prompt selection for zero-shot CLIP-based anomaly detection (Sec. IV-C).

II. Related Work

Unsupervised anomaly detection. To detect anomalies without supervised labels, recent advances in unsupervised anomaly detection use autoencoders [8], [54], deep one-class classification [60], [61], transfer learning [13], [57], [59], [63], and self-supervised learning [3], [26], [55], [67]. For accurate detection, these architectures and training frameworks depend on various hyper-parameters [5], [21], [23], but selecting hyper-parameters often requires labeled validation data which we assume is not unavailable. Similarly, semi-supervised anomaly detection methods [12], [22], [37], [69] also assume access to a training set with labeled anomalies.

Submitted 13 November 2024; revised 28 January 2025 and 26 March 2025; accepted 12 April 2025. *(Corresponding Author: Clement Fung).*

C. Fung. is with the Software and Societal Systems Department, Carnegie Mellon University, Pittsburgh, PA 15213 (email: clementf@andrew.cmu.edu). A. Li is with the Department of Computer Science, University of California Irvine, Irvine, CA 92697 (email: aodongl1@uci.edi). C. Qiu and M. Rudolph are with the Bosch Center for Artificial Intelligence, Pittsburgh, PA 15222 (email: Chen.Qiu@us.bosch.com, Maja.Rudolph@us.bosch.com)

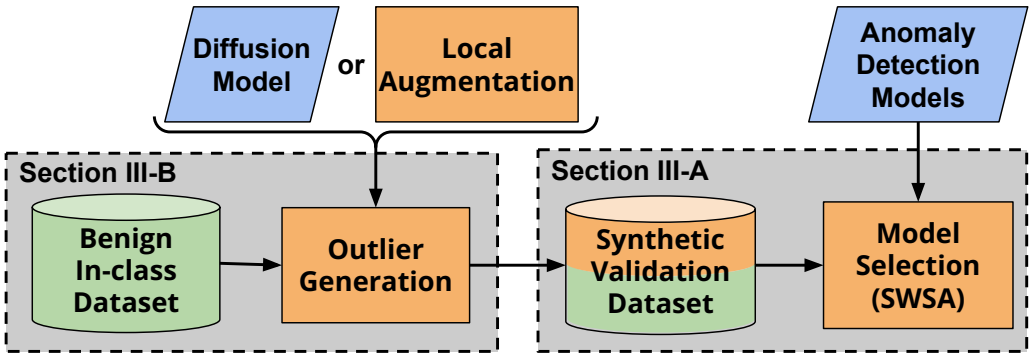


Fig. 1: We propose and compare two methods for generating synthetic anomalies, as described in Sec. III-B: image-guided generation with a diffusion model and local augmentation. By combining real normal images with synthetic anomalies, we create a synthetic validation set which is then used for model selection, as described in Sec. III-A. Components in blue are frozen, components in green are real data, and components in orange are methods implemented in this work.

Anomaly detection with foundation models. Foundation models are pre-trained on massive datasets to learn rich semantic image features and can be used for new anomaly detection tasks without additional training. Examples include vision transformers [16] (ViT) and residual networks [24] (ResNet) pre-trained on the ImageNet dataset [14]. Vision-language models, such as CLIP [58], are another powerful class of foundation model used for anomaly detection. Prior work applies CLIP to new anomaly detection [41], [42], [80] or anomaly segmentation tasks [30], [79] without training data. However, CLIP relies on the choice of text prompts; prior work learns the best-performing prompts from data [18], [30], [40], [41], [80], but we assume that this additional training or validation data is not available.

Alternative model selection strategies. Various prior works propose strategies for model selection in new, untested domains. Prior work uses internal metrics computed from predicted anomaly scores on unlabeled data [44], [45], [46], [50], but only focus on tabular data. Meta-training is another approach for anomaly detector selection [64], [77], [78], but requires several relevant labeled benchmark datasets. Finally, other prior work explores model selection with limited data but focuses on contexts that differ from ours, such model selection during training [65] or for NLP tasks [75].

Generating synthetic images from new distributions. Generative adversarial networks (GANs) [32], [33] and diffusion models [27], [68] are state-of-the-art models for image synthesis [15]. These models are commonly trained to generate images within the training data distribution, rather than the anomalous images needed for model validation. Prior work uses text prompts and CLIP to guide image synthesis towards a new distribution of interest (e.g. "cat with glasses") [20], [34], [35], [36], [49], [70], which can be used for classifier evaluation [43] and model diagnosis [28]. However, these works rely on text prompts and therefore assume a known distribution of interest. Since our work assumes that the anomalous distribution is unknown, we use DiffStyle [29], an image-generation interpolation method that uses a pre-trained diffusion model without any training, fine-tuning, or text prompts. We find that interpolating between two normal images can preserve dominant visual features (i.e., realistic background) and introduce manipulations similar to those observed in anomalies. Finally, a variety of data augmentation methods create anomalies by directly modifying normal images. These methods crop, rotate, paste, and interpolate images to create anomalous patterns [6], [39], [73].

III. Method

In settings with limited training and validation data, two commonly proposed approaches to generate synthetic data use data augmentation [73] or model-based generation [31]. These approaches work well with limited data but require domain-specific adaptations when applied to new domains. In our work, we assume that no training, fine-tuning, or domain-specific methods are used; we evaluate these two approaches for generating synthetic anomalies, to create a *synthetic validation dataset* for model selection. We call our framework SWSA. We describe how synthetic anomalies are used for SWSA in Sec. III-A and our methods for synthetic anomaly generation in Sec. III-B. Fig. 1 shows the overall process used in SWSA.

A. Model Selection with Synthetic Anomalies

Although labeled validation data is often absent when deploying anomaly detection methods, normal data is often available. For this reason, we assume access to a set of normal samples which we call the *support set* X_{support}, which is used to construct a synthetic validation set and perform model selection in the following steps:

Step 1: Partitioning the support set. We randomly partition X_{support} into seed images X_{seed} and normal validation images X_{in}. X_{seed} is used for anomaly generation, and X_{in} is held out for evaluation.

Step 2: Generating synthetic anomalies. We process X_{seed} with DiffStyle [29] or CutPaste [39] to generate synthetic anomalies $\tilde{X}_{\text{out}}$. Additional details are provided in Sec. III-B.

Step 3: Mixing the synthetic validation set. We combine X_{in} and $\tilde{X}_{\text{out}}$ to produce a labeled synthetic validation set,

$$\mathcal{D} = \{(x, 1) | x \in \tilde{X}_{\text{out}}\} \cup \{(x, 0) | x \in X_{\text{in}}\}, \tag{1}$$

where a label of 1 indicates an anomaly and a label of 0 indicates a normal image.

Step 4: Evaluating candidate models. We evaluate candidate models by their detection performance on the synthetic validation set $\mathcal{D}$. We use AUROC, the area under the receiving operator characteristic curve, which is typically used to evaluate anomaly detection models [17].

B. Generating Synthetic Anomalies

We propose two methods for generating synthetic anomalies: an augmentation-based method and a diffusion-based method. Both methods do not require any training or additional data beyond X_{seed}.

Augmentation-based. CutPaste [39] is a data augmentation method used to generate data for training unsupervised anomaly detection models. To modify an image with CutPaste, one randomly crops a region of an image and pastes it onto a different location. In our work, we propose a different use case for CutPaste: we modify seed images from X_{seed} to generate synthetic anomalies for model selection.

Diffusion-based. We use DiffStyle [29], diffusion-based style transfer, to generate synthetic anomalies with a pretrained DDIM. We first equally divide X_{seed} into style images X_{style} and content images X_{content}. DiffStyle takes any style-content image pair $\{I^{(1)}, I^{(2)}\}$ as input and generates a new image with $I^{(2)}$'s content and $I^{(1)}$'s style. To achieve this, $I^{(1)}$ and $I^{(2)}$ are mapped into the diffusion model's latent space through the forward diffusion process to produce latent vectors $x_T^{(1)}$ and $x_T^{(2)}$. We refer to the h-space (i.e., the inner-most layer of the UNet) of $x_T^{(1)}$ and $x_T^{(2)}$ as $h^{(1)}$ and $h^{(2)}$ respectively. Prior work has shown that h-space is a semantic space for images and can be manipulated during the reverse diffusion process [36].

Given two latent vectors $h^{(1)}$ and $h^{(2)}$, we perform a linear interpolation: $h^{(\text{gen})} = (1-\gamma)h^{(1)} + \gamma h^{(2)}$ where γ represents the relative weight of the content image. We then perform the asymmetric reverse diffusion process using $x_T^{(1)}$, replacing the h-space with $h^{(\text{gen})}$:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t(\epsilon_t^\theta(x_T^{(1)}|h^{(\text{gen})})) + \mathbf{D}_t(\epsilon_t^\theta(x_T^{(1)})). \quad (2)$$

We then save the final output x_0 as a synthetic anomaly. To generate our full set of synthetic anomalies $\tilde{X}_{\text{out}}$, we use all possibilities of $(I^{(1)}, I^{(2)})$ in the cross product of X_{style} and X_{content}. Although prior results [35], [36] suggest that a domain-specific diffusion model is required to generate high-quality images, we find that using one common diffusion model can be effective for SWSA. For all datasets, we use the same diffusion model pre-trained on the ImageNet dataset from prior work [15]. We discuss the hyper-parameters used for DiffStyle in Appendix A.

Fig. 2 shows examples of our diffusion-based synthetic anomalies; each synthetic anomaly (red) is interpolated from a style image (green) and a content image (cyan). We find that these images maintain the backgrounds and textures found in normal images, but introduce semantic differences that are similar to anomalies. We propose that the best-performing models at detecting these synthetic anomalies may indeed be the best-performing models at detecting real anomalies.

IV. Empirical Study

We evaluate our synthetic anomalies for model selection by investigating whether SWSA selects similar models and configurations as selection with real data. We evaluate on both natural and industrial image domains. We first describe the datasets, anomaly detection tasks, and anomaly generation methodology in Sec. IV-A. Next, we demonstrate two use cases of SWSA: model selection and CLIP prompt selection—we find using SWSA produces the best model-selection performance in seven of eight settings (Sec. IV-B) and produces the best prompt-selection performance in six of eight settings (Sec. IV-C), without any access to the real validation data.

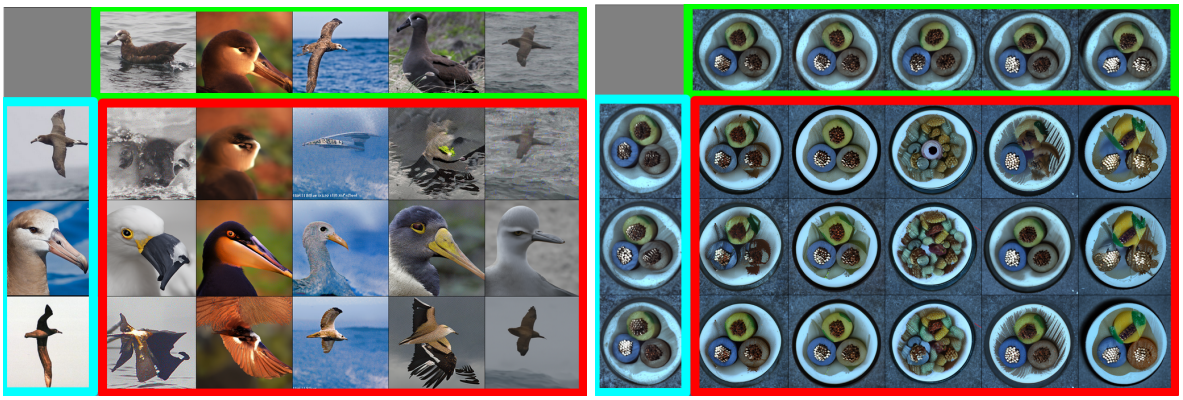
A. Experimental Setup

We present our experimental setup for evaluating anomaly detection models with synthetic validation data. We evaluate on anomaly detection tasks from four datasets and two anomaly-detection settings (i.e., one-vs-closest and one-vs-rest). We investigate how well results on synthetic validation data correspond to results with real validation data; we estimate detection performance, perform model selection, and select hyper-parameters (e.g., CLIP prompts).

Datasets. We evaluate with four frequently-used image datasets: Caltech-UCSD Birds (CUB) [71], Oxford Flowers [52], MVTec Anomaly Detection (MVTec-AD) [4], and Visual Anomaly Detection (VisA) [81]. CUB and Flowers are multi-class datasets of 200 bird species and 102 flower species respectively. MVTec-AD and VisA are datasets of multiple industrial products (15 in MVTec-AD, 12 in VisA); for each product, the training set contains images of defect-free products, and the test set contains labeled images of both defect-free and defective products.

Anomaly detection tasks. For CUB and Flowers, we create anomaly detection tasks by treating each class as normal in a one-vs-rest setting. We also adopt the one-vs-closest setting used by Mirzaei et al. [48] to simulate more difficult anomaly detection tasks. Specifically, after individually selecting each class as the inlier class, we consider each out-class individually and report the class with the worst performance. For MVTec-AD and VisA, we predict if an image is of a defective product. Each product includes images with different detect types; we consider all defect types as a single anomalous class in the one-vs-rest setting and the worst-performing defect type in the one-vs-closest setting. For all tasks, we use images from the in-class training subset as X_{support} and images from the relevant in-class and out-class validation subsets as the real validation set. In total, we evaluate with 329 anomaly detection tasks: 15 from MVTec-AD, 12 from VisA, 200 from CUB, and 102 from Flowers.

Generating synthetic anomalies. For all 329 anomaly detection tasks, we generate synthetic anomalies by drawing X_{support} from the training set of the in-class distribution only. We generate the same number of images with the diffusion-based and CutPaste-based methods: for CUB, VisA, and MVTec-AD, we sample 20 images for X_{seed} and generate 100 synthetic anomalies with each method; for Flowers, only 10 images are included in the training set for each class, so



(a) CUB class 1 (“Black Footed Albatross”) (b) MVTec-AD cable

Fig. 2: Synthetic anomalies are generated with our diffusion-based method for CUB class 1 (left) and MVTec-AD “cable” (right). Each generated image (in red) is produced from a “style” image (on top, in green) and a “content” image (on left, in cyan). All style and content images are drawn from the support set; no validation data or images from other classes are used. The generated images (in red) are then labeled as anomalies in our synthetic validation sets, which are then used for evaluating candidate anomaly detection models in SWSA.

we generate 25 synthetic anomalies with each method. Fig. 2 shows 15 examples of generated synthetic anomalies for a single CUB class (left) and MVTec-AD product (right).

B. Model Selection with Synthetic Data

We first demonstrate SWSA for model selection. Given a set of candidate models, we show that SWSA can select the true best-performing model.

Candidate anomaly detection models. We experiment across five pre-trained ResNet models (ResNet-152, ResNet-101, ResNet-50, ResNet-34, ResNet-18) and five pre-trained Vision Transformers (ViT-H-14, ViT-L-32, ViT-L-16, ViT-B-32, ViT-B-16). For all models, we use the pre-trained ImageNet weights from prior work [16], [24].

Deep-nearest-neighbor anomaly detection. To perform anomaly detection, we use the nearest-neighbor-based method of Bergman et al. [2]. We use the values of a candidate model's penultimate layer as the output of a feature extractor F and process X_{support} with F to establish a feature bank Z:

$$z_s = F(x_s), \forall x_s \in X_{\text{support}} \tag{3}$$

To perform anomaly detection on an input example d, we use the Euclidean distance between $F(d)$ and its k-nearest neighbors in Z as an anomaly score s:

$$s = \sum_{z_s \in Z_k(d)} ||F(d) - z_s||_2^2 \tag{4}$$

where $Z_k(d)$ are the k-nearest neighbors to d in the feature bank. We use $k = 3$, as suggested by Bergman et al. [2].

Evaluation setup. For each task, we calculate the AUROC for each candidate model using the synthetic and real validation datasets. We average the respective AUROCs across all anomaly detection tasks to calculate “synthetic AUROC” and “real validation AUROC”. We then compare the synthetic AUROC and real validation AUROC to investigate if the rankings of candidate models are similar. When reporting the AUROC of SWSA, we select the model with the best synthetic AUROC and report its corresponding real validation AUROC.

As a baseline, we include SWSA using the Tiny-ImageNet dataset. Prior work uses Tiny-Imagenet for fine-tuning anomaly detection models (i.e., outlier exposure [25]), and we investigate if Tiny-Imagenet is effective for model selection; we randomly sample images from Tiny-ImageNet to generate $\tilde{X}_{\text{out}}$ of the same size: 100 images for tasks with CUB, VisA, and MVTec-AD; 25 images for tasks with Flowers.

Evaluation results. We first evaluate SWSA for model selection. The top half of Table I shows how often the best model is picked (i.e., pick rate) and the resulting AUROC for different model selection strategies. SWSA with diffusion-based anomalies or CutPaste-based anomalies often selects the best model and produces the highest AUROC for six out of eight evaluation settings, even outperforming the largest available model (ViT-H-14). In particular, SWSA with diffusion-based anomalies selects the best model the most often for all CUB and Flowers settings.

We also evaluate SWSA for model ranking, beyond selecting the best-performing model. Fig. 3 shows the synthetic and real validation AUROC for all 10 models in the one-vs-closest (top) and one-vs-rest (bottom) settings. For Flowers and CUB, SWSA is most consistent with diffusion-based anomalies in the one-vs-rest setting. We show the Kendall's Tau rank correlation coefficients between the synthetic and real validataion AUROC in Table II.[1] For most one-vs-rest anomaly detection tasks, we also find that SWSA performs well with few anomalies; we vary the number of synthetic anomalies from the full set of anomalies to as few as five, keeping the anomalies with the lowest anomaly score (i.e., the most difficult anomalies). SWSA with diffusion-based anomalies performs best on most one-vs-rest tasks with Flowers.

For MVTec-AD and VisA, unlike the datasets of natural images, SWSA is less effective; the model selection results

[1]For Tiny-Imagenet, the synthetic AUROC $\approx$ 1.0 for most cases, and the rank correlation is near zero.

TABLE I: We report (i) how often SWSA picks the best model/prompt ("pick rate") and (ii) the resulting AUROC of the selections made by SWSA. We show the AUROC when the best model/prompt is always selected (in grey) as an upper bound. For all settings, SWSA outperforms baseline strategies: using the largest model (ViT-H-14), the default prompt, or a prompt ensemble [58]. In particular, SWSA with diffusion-based anomalies is most effective for natural images (CUB and Flowers).

		CUB		Flowers		MVTec-AD		VisA	
		Pick rate	AUROC	Pick rate	AUROC	Pick rate	AUROC	Pick rate	AUROC
One-vs-Closest (model selection)	Largest model	11 / 200	0.653	43 / 102	0.956	4 / 15	**0.716**	1 / 12	0.636
	SWSA (TinyImg)	30 / 200	0.674	50 / 102	0.966	4 / 15	**0.716**	1 / 12	0.636
	SWSA (Diffusion)	**66 / 200**	0.737	**59 / 102**	**0.967**	4 / 15	0.670	0 / 12	0.643
	SWSA (CutPaste)	60 / 200	**0.743**	37 / 102	0.945	2 / 15	0.678	1 / 12	**0.674**
	Best Model	–	0.826	–	0.990	–	0.785	–	0.752
One-vs-Rest (model selection)	Largest model	32 / 200	0.982	49 / 102	**0.994**	4 / 15	**0.733**	1 / 12	0.765
	SWSA (TinyImg)	59 / 200	0.982	57 / 102	0.993	4 / 15	0.733	1 / 12	0.765
	SWSA (Diffusion)	**109 / 200**	**0.988**	**62 / 102**	**0.994**	2 / 15	0.706	0 / 12	0.764
	SWSA (CutPaste)	62 / 200	0.966	37 / 102	0.974	2 / 15	0.717	1 / 12	**0.772**
	Best Model	–	0.991	–	0.997	–	0.757	–	0.824
One-vs-Closest (prompt selection)	Default Prompt	5 / 200	0.571	1 / 102	0.697	2 / 15	0.741	0 / 12	0.596
	Prompt Ensemble	–	0.577	–	0.708	0 / 15	0.728	0 / 12	0.596
	SWSA (TinyImg)	34 / 200	0.582	18 / 102	0.718	2 / 15	0.760	1 / 12	**0.612**
	SWSA (Diffusion)	**46 / 200**	**0.590**	**38 / 102**	**0.729**	2 / 15	0.725	2 / 12	0.596
	SWSA (CutPaste)	34 / 200	0.585	18 / 102	0.718	3 / 15	**0.763**	2 / 12	0.604
	Best Prompt	–	0.625	–	0.759	–	0.845	–	0.702
One-vs-Rest (prompt selection)	Default Prompt	0 / 200	0.971	1 / 102	0.959	2 / 15	0.752	0 / 12	0.724
	Prompt Ensemble	–	0.972	–	0.962	0 / 15	0.753	0 / 12	**0.747**
	SWSA (TinyImg)	27 / 200	0.972	13 / 102	**0.967**	5 / 15	**0.765**	1 / 12	0.745
	SWSA (Diffusion)	**64 / 200**	**0.973**	**38 / 102**	**0.967**	1 / 15	0.728	2 / 12	0.724
	SWSA (CutPaste)	27 / 200	0.972	22 / 102	0.963	1 / 15	0.746	2 / 12	0.730
	Best Prompt	–	0.976	–	0.971	–	0.786	–	0.801

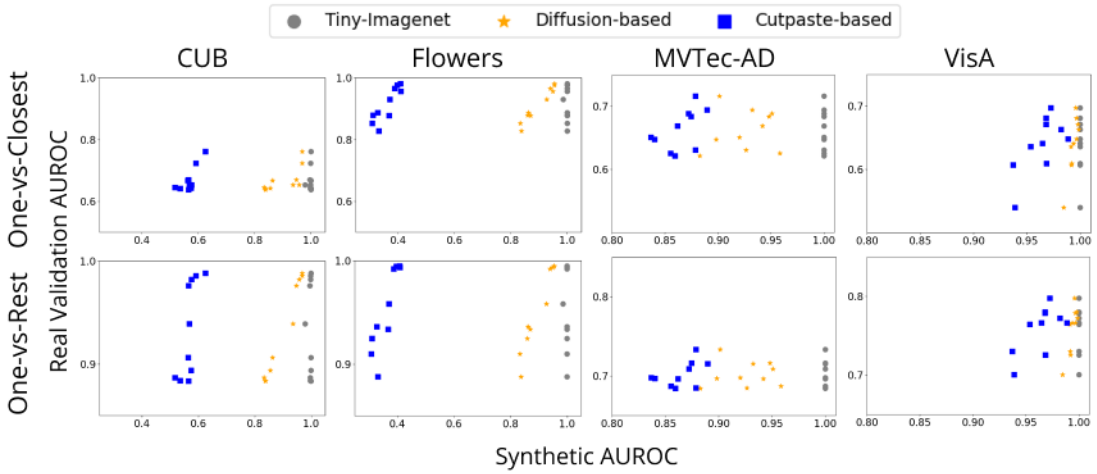


Fig. 3: We compare the rankings of real and synthetic validation AUROC for all models across three types of synthetic anomalies: Tiny-Imagenet, diffusion-based anomalies, and Cutpaste-based anomalies. Ideally, the ranking of models with synthetic data (along the x-axis) should match the ranking of models with real data (along the y-axis). SWSA performs best when ranking models with diffusion-based anomalies in the one-vs-rest anomaly detection setting on CUB and Flowers. We provide Kendall's Tau rank correlation values for these results in Table II.

are less consistent and the correlation from model rankings are not statistically significant. These anomalies come from fine-grained industrial defects and are generally more difficult to detect; we provide additional analysis in Sec. IV-D.

C. CLIP Prompt Selection with Synthetic Data

In this section, we show that SWSA can be used to select the best-performing prompts for CLIP-based anomaly detection.

Zero-shot anomaly detection with CLIP. We use CLIP to perform anomaly detection using a technique from prior work: given an input image, we input two text prompts to CLIP—one for normal data, one for anomalies—and predict based on the prompt that produces a higher similarity to the image [58]. We use the same backbone ("ViT-B-16-plus-240") and data transformations as in prior work [30]. In our candidate text prompts, we assume that the name of

TABLE II: To evaluate SWSA for model ranking, we calculate the Kendall's Tau rank correlation between the rankings found with real and synthetic validation datasets for diffusion-based and CutPaste-based anomalies. Since we perform repeated tests on the same data, we apply Bonferroni correction; cases with a statistically significant rank correlation *($p < 5.56e\text{-}3$)* are **bolded**. We find that SWSA is particularly effective with diffusion-based anomalies for (i) one-vs-rest tasks on CUB and Flowers, and (ii) one-vs-closest tasks on Flowers.

		# of synthetic anomalies		
		All anomalies	10	5
One-vs-Closest	Diffusion (CUB)	0.644 (*p=9.14e-3*)	0.600 (*p=1.67e-2*)	0.600 (*p=1.67e-2*)
	CutPaste (CUB)	0.377 (*p=1.55e-1*)	0.511 (*p=4.66e-2*)	0.555 (*p=2.86e-2*)
	Diffusion (Flowers)	**0.777 (*p=9.46e-4*)**	**0.777 (*p=9.46e-4*)**	**0.822 (*p=3.57e-4*)**
	CutPaste (Flowers)	0.644 (*p=9.14e-3*)	0.466 (*p=7.25e-2*)	0.244 (*p=3.81e-1*)
One-vs-Rest	Diffusion (CUB)	**0.866 (*p=1.15e-4*)**	**0.822 (*p=3.57e-4*)**	**0.822 (*p=3.57e-4*)**
	CutPaste (CUB)	0.600 (*p=1.67e-2*)	**0.733 (*p=2.21e-3*)**	**0.688 (*p=4.68e-3*)**
	Diffusion (Flowers)	**0.866 (*p=1.15e-4*)**	**0.866 (*p=1.15e-4*)**	**0.911 (*p=2.97e-5*)**
	CutPaste (Flowers)	**0.733 (*p=2.21e-3*)**	0.555 (*p=2.86e-2*)	0.333 (*p=2.16e-1*)

the normal class is known. For CUB and Flowers, we use "`some`" to describe anomalies; for example we compare "`a photo of a red cardinal`" to "`a photo of some bird`". For MVTec-AD and VisA, we use "`with defect`" to describe anomalies; for example, we compare "`a photo of a transistor`" to "`a photo of a transistor with defect`". We select from a set of ten prompts used in prior work [30], described in Appendix B.

Evaluation setup. We compare SWSA to two baselines: (i) the default prompt template (i.e., "`a photo of a [class name] bird`" vs "`a photo of some bird`") and (ii) a full prompt ensemble as proposed by Radford et al. [58] and evaluated in prior work [30], [80]. For each selection strategy, we select the prompt with the best synthetic AUROC and report (i) how often the selection matches the best prompt on real validation data and (ii) the resulting AUROC when using the selected prompt.

Evaluation results. We evaluate SWSA for prompt selection in the one-vs-closest and one-vs-rest settings for all 329 anomaly detection tasks. In the bottom half of Table I, we report how often each strategy selects the best prompt and each strategy's resulting AUROC. SWSA with diffusion-based anomalies performs best for Flowers and CUB by selecting the best prompt the most often and producing the highest AUROC for all settings, outperforming the popular prompt ensemble.

For MVTec-AD and VisA, SWSA performs best with CutPaste-based anomalies, outperforming the prompt ensemble in one-vs-closest settings. We find that although the commonly proposed prompt ensemble is most effective in general, it does not always perform best; SWSA outperforms the prompt ensemble in particularly difficult settings (i.e., worst-case anomaly detection tasks). Overall, SWSA can be used to select the best prompts for CLIP-based anomaly detection for tasks of varying domain and difficulty.

D. Analysis of SWSA

In this section, we use four representative one-vs-closest anomaly detection tasks to analyze SWSA's performance on different types of synthetic anomalies. To highlight different failure cases, we analyze the task from each dataset with the lowest baseline AUROC.

TABLE III: For the four anomaly detection tasks described in Sec. IV-D, we compute the total variation between the real and synthetic validation datasets. We find that no setting provides a tight bound and we cannot provide strong guarantees of rank preservation with SWSA.

Real vs:	Flowers	CUB	MVTec-AD	VisA
Diffusion-based	0.698	1.257	1.143	0.934
Cutpaste-based	0.697	1.042	0.534	0.804

Theoretical analysis. We first perform a theoretical analysis of SWSA by computing the total variation between real and synthetic validation sets to determine if a tight bound exists for model rank preservation. Shoshan et al. [65] study model selection with synthetic data for binary classification; they show that the total variation distance between a synthetic validation set and true data provides an upper bound on the empirical risk difference between any two classifiers. We follow the method of Sajjadi et al. [62] to compute the total variation between datasets. For each dataset, we compare $D_1 = X_{\text{support}} \cup X_{\text{out}}$ and $D_2 = X_{\text{support}} \cup \tilde{X_{\text{out}}}$, where $\tilde{X_{\text{out}}}$ are our diffusion-based or CutPaste-based synthetic anomalies.

Table III shows the total variation for each dataset and type of synthetic anomaly. We find that neither Cutpaste-based or diffusion-based anomalies provide a tight bound for synthetic validation sets (the empirical risk difference is greater than 0.5). Although we ultimately cannot provide theoretical guarantees, our empirical results in Table I show that SWSA often selects models and prompts that match the selections made with real validation data. An analysis of SWSA that provides stronger guarantees is left as future work.

Qualitative analysis. We next perform a qualitative visual analysis of our representative anomaly detection tasks. Fig. 4 (top) shows the t-SNE visualization of the embeddings from the ViT-B-16 model for each task. We plot the embeddings of (i) real normal images, (ii) real abnormal images, (iii) diffusion-generated anomalies, and (iv) CutPaste-generated anomalies. Real anomalies are not available for SWSA in practice but we use their embeddings for illustration. We find that, when anomalies come from natural variations (e.g., differ-

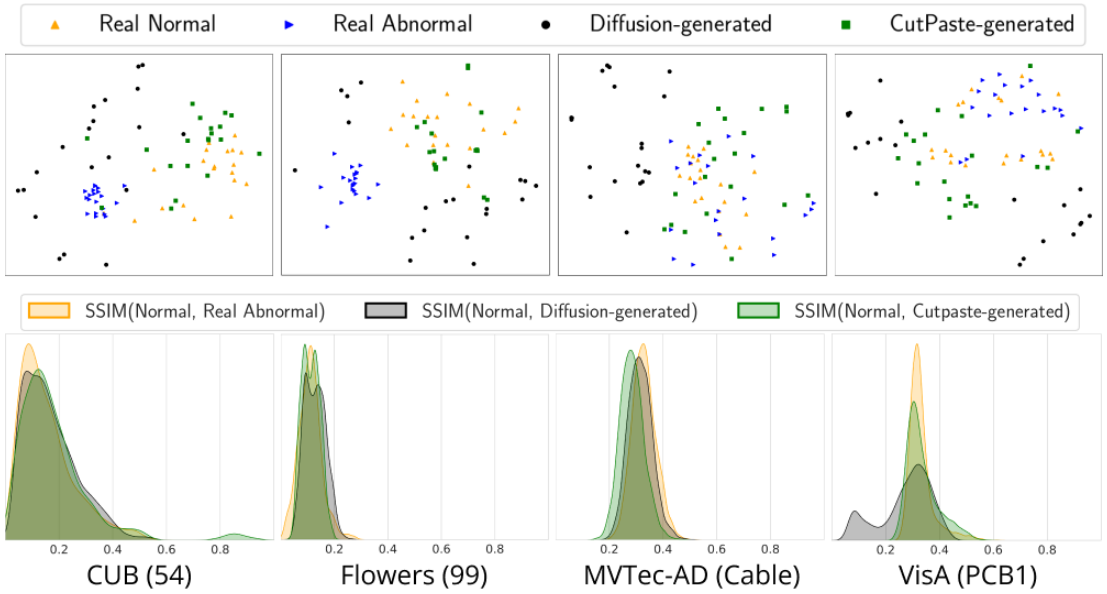


Fig. 4: For four representative anomaly detection tasks, we compare real one-vs-closest anomalies (blue triangle), diffusion-generated anomalies (black circle), and Cutpaste-generated anomalies (green square). On top, we show t-SNE visualizations of the ViT-B-16 embeddings for each image. On bottom, we plot a distribution of the SSIM scores between each pair of anomaly types. Overall, when anomalies come from natural variations between classes (CUB and Flowers), they are further from the normal class and are better represented by diffusion-based anomalies. When anomalies come from fine-grained changes (MVTec-AD and VisA), they are closer to normal images and are better represented by Cutpaste-based anomalies.

ent flower species), their embeddings are further from normal data, similar to diffusion-based anomalies. Conversely, when anomalies come from fine-grained changes (e.g., defective pin on a chip), their embeddings are closer to normal data, similar to Cutpaste-based anomalies.

We also compare the structural similarity (SSIM) [72] for different types of anomalies. We calculate the SSIM between real normal images and each type of anomalies; a higher SSIM indicates that the pairs of images are more structurally similar. Fig. 4 (bottom) shows the kernel density estimate of the distribution of SSIM scores for each type of anomaly on each representative anomaly detection task. Although SSIM is only a heuristic for the visual difference between images, we note that SWSA's performance depends on how well diffusion-generated or Cutpaste-based anomalies can represent the distances typically observed between normal images and anomalies; both CUB and Flowers show higher distributional overlaps with CutPaste-based and diffusion-based anomalies. Finally, we find that for MVTec-AD and VisA, the distributions of real anomalies overlap more closely with the normal data distribution; their t-SNE embeddings are closer and their SSIM scores are higher. This makes detecting these anomalies with foundation models more difficult and suggests why SWSA is worse at estimating their performance.

V. Limitations and Future Work

Integrating SWSA in real-world settings and other domains. In real-world settings, we assume that practitioners have access to a set of candidate anomaly detection models, for which model selection is needed. In practice, integrating SWSA only requires a small set of normal images (for the support set) and a pre-trained, public diffusion model (for image synthesis).

Although the computational and data requirements for SWSA are low, we anticipate challenges with domain transfer when integrating SWSA to new domains. We found that SWSA performed most poorly on datasets of industrial defects; in general, anomaly detection for these tasks is difficult with pre-trained models [42], [48] and requires fine-tuning or domain-specific prompt tuning [30], both of which are out of scope in our setting.

Our work uses pre-trained models on ImageNet, and we expect that integrating SWSA for domains not well represented by ImageNet will require new foundation models, such as (i) those trained for industrial products similar to those found in MVTec-AD and VisA or (ii) for non-image domains, such as text [76] and time-series data [74]. As these models become available, applying SWSA to such domains will become an intriguing area for future work.

Security considerations for SWSA. SWSA relies on pre-trained diffusion models for anomaly generation. Prior work has shown that attackers can inject backdoors [7], [10], [11] into diffusion models. An attack could inject backdoors into a diffusion model which, when used in SWSA, could mislead a victim into selecting a sub-optimal anomaly detection model, enabling other attacks. Neuron pruning and latent clipping have been proposed as countermeasures to these attacks [10], and exploring how these attacks and defenses affect SWSA is interesting future work.

Next opportunities for improving SWSA. While our results show initial promise for SWSA, especially on anomaly detection tasks with natural images, we believe that there

are still promising opportunities for future work to improve SWSA further and apply our synthetic anomalies for other related tasks. First, our work evaluates data augmentation and diffusion-based methods independently; we propose that hybrid generation techniques which combine techniques from diffusion-based style transfer, image interpolation, and image modification may be effective in generating anomalies for more difficult tasks. Second, we suggest that SWSA may be effective in meta-learning and active-learning settings; in these settings, our synthetic anomalies could be used both for selection and for fine-tuning candidate anomaly detection models, similar to outlier exposure [25], [38], [42], [53], [56], which explores fine-tuning of anomaly detection models with auxiliary data. Finally, our work only performs a minimal amount of prompt design since we assume that the anomalous domain is completely unknown. In contrast, if we can make some minimal assumptions about anomalies, we can use this information to further engineer and optimize the candidate prompt templates used for SWSA, improving the performance of CLIP-based anomaly detection.

VI. Conclusion

In this work, we propose and evaluate SWSA: an approach to select image-based anomaly detection models without validation data or domain-specific methods. We use a general-purpose diffusion model to generate synthetic anomalies using only a small support set of in-class examples, without requiring any model training or fine-tuning. We present an empirical study which shows that SWSA can be used to select image-based anomaly detection models and to select prompts for zero-shot CLIP-based anomaly detection. SWSA can outperform baseline selection strategies, such as using the largest model or a prompt ensemble.

References

[1] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55(C):278–288, 2016.

[2] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020.

[3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[5] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30, 2016.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 2002.

[7] Weixin Chen, Dawn Song, and Bo Li. TrojDiff: Trojan attacks on diffusion models with diverse targets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In *MIDL Conference Book*, 2018.

[9] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[11] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. VillanDiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 2023.

[12] Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern, and Andrew Emmott. Incorporating expert feedback into active anomaly discovery. In *International Conference on Data Mining*, 2016.

[13] Lucas Deecke, Lukas Ruff, Robert A Vandermeulen, and Hakan Bilen. Transfer-based semantic anomaly detection. In *International Conference on Machine Learning*, 2021.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*, 2015.

[18] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model CLIP. In *AAAI conference on artificial intelligence*, 2022.

[19] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys*, 54(7), 2021.

[20] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[21] Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4), 2016.

[22] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46, 2013.

[23] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 2022.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[25] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[26] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 2019.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *International Conference on Learning Representations*, 2023.

[29] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[30] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[31] Yanmei Jiang, Xiaoyuan Ma, and Xiong Li. Towards virtual sample generation with various data conditions: A comprehensive review. *Information Fusion*, 117:102874, 2025.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image

editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[35] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[36] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*, 2023.

[37] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Stephan Mandt, and Maja Rudolph. Deep anomaly detection under labeling budget constraints. In *International Conference on Machine Learning*, 2023.

[38] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection without foundation models. *arXiv preprint arXiv:2302.07849*, 2023.

[39] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[40] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[41] Yiting Li, Adam Goodge David, Fayao Liu, and Chuan-Sheng Foo. PromptAD: Zero-shot anomaly detection using text prompts. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[42] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus Robert Muller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022.

[43] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[44] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter*, 25(1), 2023.

[45] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4), 2020.

[46] Henrique O Marques, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. On the internal evaluation of unsupervised outlier detection. In *International conference on scientific and statistical database management*, 2015.

[47] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. Kitsune: an ensemble of autoencoders for online network intrusion detection. In *Network and Distributed System Security Symposium*, 2018.

[48] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees G. M. Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it until you make it: Towards accurate near-distribution novelty detection. In *International Conference on Learning Representations*, 2023.

[49] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[50] Thanh Trung Nguyen, Uy Quang Nguyen, et al. An evaluation method for unsupervised anomaly detection algorithms. *Journal of Computer Science and Cybernetics*, 32(3), 2016.

[51] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

[52] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[53] Lorenzo Perini, Maja Rudolph, Sabrina Schmedding, and Chen Qiu. Uncertainty-aware evaluation of auxiliary anomalies with the expected anomaly posterior. *Transactions on Machine Learning Research*, 2025.

[54] Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. Acoustic novelty detection with adversarial autoencoders. In *2017 International Joint Conference on Neural Networks*, 2017.

[55] Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. Raising the bar in graph-level anomaly detection. In *International Joint Conference on Artificial Intelligence*, 2022.

[56] Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *International Conference on Machine Learning*, 2022.

[57] Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In *International Conference on Machine Learning*, 2021.

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[59] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[60] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, 2018.

[61] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2019.

[62] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 2018.

[63] Tim Schneider, Chen Qiu, Marius Kloft, Decky Aspandi Latif, Steffen Staab, Stephan Mandt, and Maja Rudolph. Detecting anomalies within time series using local neural transformations. *arXiv preprint arXiv:2202.03944*, 2022.

[64] David Schubert, Pritha Gupta, and Marcel Wever. Meta-learning for automated selection of anomaly detectors for semi-supervised datasets. In *International Symposium on Intelligent Data Analysis*, 2023.

[65] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, Matan Fintz, and Gerard Medioni. Synthetic data for model selection. In *International Conference on Machine Learning*, 2023.

[66] Hossein Siadati and Nasir Memon. Detecting structurally anomalous logins within enterprise networks. In *ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[67] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, 2020.

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[69] Holger Trittenbach, Adrian Englhardt, and Klemens Böhm. An overview and a benchmark of active learning for outlier detection with one-class classifiers. *Expert Systems with Applications*, 168, 2021.

[70] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[71] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[72] Zhou Wang. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004.

[73] Chaobin Xu, Wei Li, Xiaohui Cui, Zhenyu Wang, Fengling Zheng, Xiaowu Zhang, and Bin Chen. Scarcity-GAN: Scarce data augmentation for defect detection via generative adversarial nets. *Neurocomputing*, 566(C), 2024.

[74] Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*, 2024.

[75] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. Test accuracy vs. generalization gap: Model selection in NLP without accessing training or testing data. In *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[76] Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024.

[77] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. *Advances in Neural Information Processing Systems*, 2021.

[78] Yue Zhao, Sean Zhang, and Leman Akoglu. Toward unsupervised outlier model selection. In *IEEE International Conference on Data Mining*, 2022.

[79] Chong Zhou, Chen Change Loy, and Bo Dai. DenseCLIP: Extract free dense labels from CLIP. *arXiv preprint arXiv:2112.01071*, 2021.

[80] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2024.

[81] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 2022.

APPENDIX

A. Hyper-parameters for Diffusion-based Image Synthesis

We first discuss the impact of γ on diffusion-based image generation, which controls the interpolation strength, shown in Fig. 5. γ represents the relative strength of the content image and should be high enough to introduce anomalous patterns. γ also affects the SSIM from the original style image, which suggests that if some assumptions about true anomalies are known, image similarity scores can be used as a heuristic for selecting diffusion hyper-parameters, as is done in prior work [29]. For our anomaly detection tasks, we find that using $\gamma = 0.7$ generates images that differ significantly from the original style images, while preserving elements near the in-class distribution.

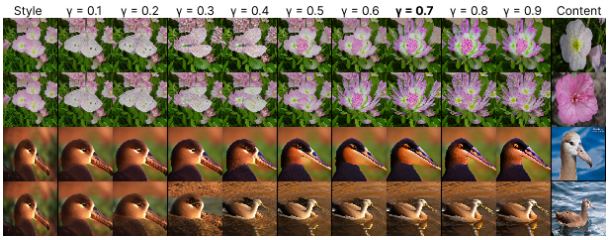


Fig. 5: We compare values of γ used in the diffusion-based generation. A higher value of γ corresponds to a stronger weight for content in the resulting interpolation.

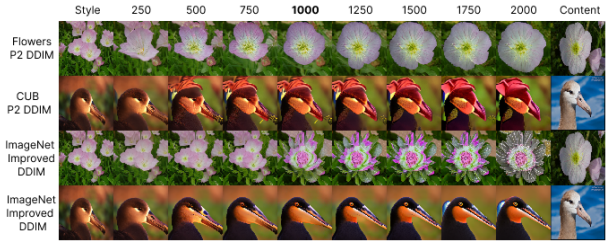


Fig. 6: We compare the diffusion model types and the number of iterations used for diffusion-based generation. Using a higher number of iterations and the improved ImageNet diffusion model results in larger changes to the normal images.

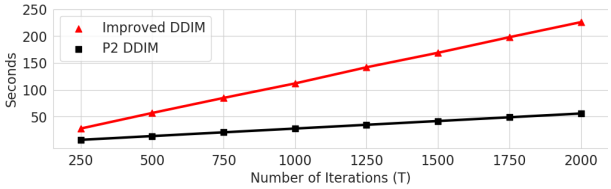


Fig. 7: We show the time taken to generate a single image with T iterations using (i) the improved ImageNet diffusion model and (ii) P2 diffusion model.

We next discuss the impact of T (the number of iterations) and the type of diffusion model. In our work, we use $T = 1000$ during the reverse DDIM process of the improved ImageNet diffusion model [51], which takes approximately 100 seconds with a single RTX 3090 GPU. Thus, it takes approximately 2.8 hours to generate our 100-image synthetic validation dataset. To speed up the diffusion process, one can use alternative types of diffusion model or reduce T; we compare the improved ImageNet model to the perception prioritized (P2) diffusion model from prior work [9], which provides weights trained on the CUB and Flowers datasets directly. Fig. 7 shows the average time taken to generate a single image using T iterations for a given diffusion model, we find that P2 diffusion models are faster; with only 250 iterations on the P2 model, generating our entire synthetic validation dataset of 100 images would take under 12 minutes. Fig. 6 shows the results of these generations; using the improved ImageNet diffusion model and a higher number of iterations introduces more disruptions to the normal images. Determining the optimal tradeoff between computation cost and image quality is left as future work.

B. Prompt Templates for CLIP-based Anomaly Detection

For our experiments in Sec. IV-C, we evaluated across a set of ten candidate prompt templates, shown below. For the results shown in Table I, "default prompt" is the first prompt, and "prompt ensemble" is the average across all prompts. For CUB and Flowers, only the term "bird" or "flower" is used in the template. For MVTec-AD and VisA, we only perform mild class-name cleaning: we remove trailing numbers from class names and fully write all acronyms (e.g., "PCB1" is written as "printed circuit board"). Unlike prior work [30], we do not perform any other class-specific modifications.

```
% CLIP Templates for Flowers
'a photo of [a {} flower, some flower]'
'a cropped photo of [a {} flower, some flower]'
'a dark photo of [a {} flower, some flower]'
'a photo of [a {} flower, some flower] for inspection'
'a photo of [a {} flower, some flower] for viewing']'
'a bright photo of [a {} flower, some flower]'
'a close-up photo of [a {} flower, some flower]'
'a blurry photo of [a {} flower, some flower]'
'a photo of a small [{} flower, some flower]'
'a photo of a large [{} flower, some flower]'

% CLIP Templates for CUB
'a photo of [a {} bird, some bird]',
'a cropped photo of [a {} bird, some bird]'
'a dark photo of [a {} bird, some bird]'
'a photo of [a {} bird, some bird] for inspection'
'a photo of [a {} bird, some bird] for viewing'
'a bright photo of [a {} bird, some bird]'
'a close-up photo of [a {} bird, some bird]'
'a blurry photo of [a {} bird, some bird]'
'a photo of a small [{} bird, some bird]'
'a photo of a large [{} bird, some bird]'

% CLIP Templates for MVTec-AD and VisA
'a photo of [a {}, a {} with defect]'
'a cropped photo of [a {}, a {} with defect]'
'a dark photo of [a {}, a {} with defect]'
'a photo of [a {}, a {} with defect] for inspection'
'a photo of [a {}, a {} with defect] for viewing'
'a bright photo of [a {}, a {} with defect]'
'a close-up photo of [a {}, a {} with defect]'
'a blurry photo of [a {}, a {} with defect]'
'a photo of a small [{}, {} with defect]'
'a photo of a large [{}, {} with defect]'
```