Data-centric Anomaly Detection with Diffusion Models

Sheldon Liu Gordon Wang Lei Liu Xuefeng Liu

Abstract—Anomaly detection, also referred to as one-class classification, plays a crucial role in identifying product images that deviate from the expected distribution. This study introduces Data-centric Anomaly Detection with Diffusion Models (DCADDM), presenting a systematic strategy for data collection and further diversifying the data with image generation via diffusion models. The algorithm addresses data collection challenges in real-world scenarios and points toward data augmentation with the integration of generative AI capabilities. The paper explores the generation of normal images using diffusion models. The experiments demonstrate that with 30% of the original normal image size, modeling in an unsupervised setting with state-of-the-art approaches can achieve equivalent performances. With the addition of generated images via diffusion models (10% equivalence of the original dataset size), the proposed algorithm achieves better or equivalent anomaly localization performance.

Keywords—Diffusion Models, Anomaly Detection, Data-centric, Generative AI

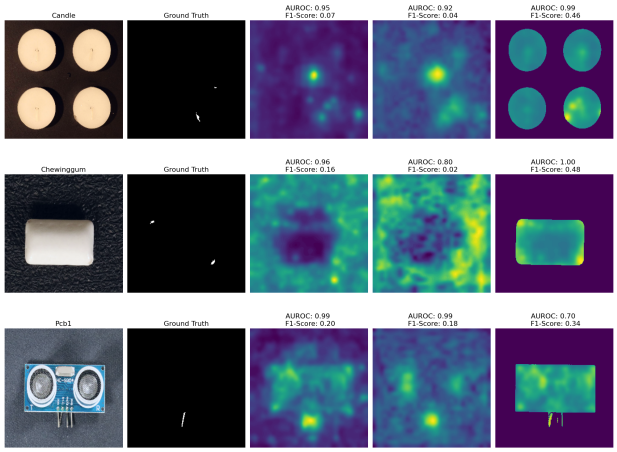


Fig. 1: Foreground-aware dataset augmentation with diffusion models

I. Introduction

Anomaly detection in product images through machine learning plays a pivotal role in ensuring product quality control. The application of imaging and computer vision algorithms has streamlined various processes, including inspection, defect reporting, retrieval of defective products, and related procedures [1], [2], [3], [4]. Humans possess an innate ability to discern expected variances in datasets and outliers after viewing only a few normal images [5]. Motivated by this observation, researchers have introduced diverse unsupervised approaches. Distribution-based algorithms like Padim [6] and Patchcore [5] create a distribution of patch-level features from normal images. Abnormal regions in testing images are identified by calculating the distance of image features to the distribution. Autoencoder-based methods are trained to minimize reconstruction loss, identifying anomalies in testing images exhibiting significant reconstruction loss [7], [8]. Alternatively, Generative Adversarial Network (GAN) based techniques [9], [10] employ adversarial training. A recent state-of-the-art method is EfficientAD [11], designed to meet real-time computer vision application requirements. This approach employs a student-teacher framework, where the student network predicts features extracted from normal images, and an anomaly is identified when the student fails to do so.

To counter the shortage of normal product images and reduce the effort of data collection, current research emphasizes data augmentation/data generation through stable diffusion and image inpainting. Traditional augmentation techniques like rotations and flips often lack diversity along critical semantic axes in the data. Stable diffusion, leveraging text-to-image generation, proves effective in generating novel visual concepts [12], [13], [14].

Image inpainting is a computer vision technique employed to restore or fill in missing or damaged regions within an image [15]. The goal is to generate a visually plausible completion of the damaged / missing areas based on the surrounding context and instruction prompts. EditorBench proposed by Google Research is providing a systematic benchmark for text-guided image inpainting [16]. Imagic, claimed to be a preferred tool by human raters, is generating a text embedding aligning with both the input image and the target text [17].

In this paper, an algorithm for Data-centric Anomaly Detection with Diffusion Models is proposed. During the diffusion processes, the diffusion model is fine-tuned with domain-specific normal images. A systematic strategy of data collection is proposed to significantly reduce the data-collection effort without sacrificing model performance. Highlighted results in three image domains are shown in Fig. 1, where the proposed algorithm is benchmarked against *Padim* and *Patchcore* measured with the metrics of AUROC and F1-Score on pixel-level. With fore-ground aware variance added to the dataset, the model is able to produce superior anomaly localization result.

The paper is organized as follows. In the Section II, a summary is provided on the works that are related with anomaly detection, and stable diffusion. The problem is then

Sheldon, Gordon, Lei and Xuefeng are scientists/managers from Generative AI Innovation Center, Amazon Web Services
Corresponding author: Sheldon Liu, email: shilong@amazon.com

formalized and the proposed approach is detailed in Section III. In Section IV, the experiments conducted in this study is included. Section V illustrates the ablation study and Section VI draws the conclusion. More qualitative results are included in Appendix A.

II. Related works

Anomaly Detection Modern methods for anomaly detection can be divided into two main paradigms, namely deep Feature-based methods and Reconstruction-based methods. Feature-based methods use CNN to extract meaningful vectors describing an entire image for anomaly detection or an image patch for anomaly localization. Self-supervised learning has been used in the past to learn image features [18], [19], [20], often solving auxiliary tasks. In anomaly detection, [21] have demonstrated that high-quality features facilitate the detection of anomalous samples. DN2 [22] has successfully employed simple ResNets [23], pretrained on Imagenet, to extract informative features. Distribution-based algorithms like Padim [6] and Patchcore create a distribution of patch-level features from normal images. Notably, Patchcore [5] incorporates a memory bank concept, achieving superior performance by incorporating a well-fitted inductive bias. This stands in contrast to SPADE [24], which utilizes a memory bank of nominal features from a pre-trained backbone network, employing distinct approaches for image- and pixel-level anomaly detection. Reconstruction-based methods first train image reconstruction models on normal images. Commonly used models include: Autoencoders [25], [26], [27], VAEs [28], GANs [29], etc

Overall, benefiting from the powerful representation capabilities of deep features, feature-based methods can achieve better performance compared to existing reconstruction-based methods. However, since it is difficult to obtain enough labeled training data in actual situations, sample imbalance will occur, and it is difficult for these methods to achieve better results.

Diffusion models In recent times, diffusion models have become a focal point of considerable interest [30], [31]. These models establish a paradigm wherein the forward process systematically introduces random noise to the data, while the reverse process reconstructs desired data samples from the noise. This framework has given rise to a diverse array of diffusion-based applications in perception, including but not limited to image generation [30], [13], [32] and image segmentation [33], [34], [35].

Within the realm of diffusion models, text-guided image inpainting stands out as a significant branch, attracting substantial scholarly attention in recent research endeavors. A notable contribution in this domain is "Paint By Word" [36], which strives to achieve a delicate balance between a) maintaining consistency between input and edited images, and b) ensuring coherence between the textual guide and the edited image. This approach has found effective application in subsequent advancements, such as DiffusionCLIP [37]. Blended Diffusion [38] adopts a unique strategy by concurrently applying CLIP-guided diffusion to the foreground (masked region) and background (context) separately, subsequently blending the outcomes through element-wise aggregation. Introducing an auto-regressive text-guided infilling technique powered by cross-modal language modeling, CogView2 [39] represents another noteworthy contribution.

DiffEdit [40] introduces a "masked mask-free" formulation where masking segmentation and masked diffusion operate in parallel for inpainting. Of particular relevance to this study are Stable Diffusion [13] and GLIDE/DALL-E2 [41], [42], both classified as diffusion models. Noteworthy advancements have also been made in mask-free text-guided image editing [17], [43]. For instance, Text2Live [43] operates on an isolated edit-layer with semantic localization, preserving context effectively but limiting extensive modifications. Prompt-to-Prompt [44] introduces potent manipulation techniques on the cross-attention in the text-conditioning module. Lastly, Imagic [17] optimizes a specialized embedding to capture the semantics of the input image, producing textually faithful edits through interpolation with the embedding of the target text. The field of text-guided image inpainting has recently witnessed significant scholarly attention, with Paint By Word [38] as a pivotal technique aiming to strike a balance between input-consistency and meaningful correlation with textual guidance. This methodology has been effectively integrated into contemporary works such as DiffusionCLIP [37].

A captivating new direction in diffusion model research is being explored, where pre-trained text-to-image diffusion models are utilized to attain detailed or nuanced manipulation of synthesis results. The innovative technique to handle the emerging challenge of subject-driven generation is presented by DreamBooth[45]. With only a few casually captured images of a subject, users are enabled to recontextualize subjects, adjust their properties, produce original art renditions, and more. In this paper, DreamBooth diffusion model are fine tuned to generate domain-specific normal images.

III. Proposed Method

A. Problem Formulation

The anomaly detection problem involves determining whether an image is normal (I^+, $y = 0$) or abnormal (I^-, $y = 1$) and providing the location of abnormal regions. In the VisA dataset, normal images $I^+_{i,c}$ are denoted for class $c \in [1, 12]$ with N_c images under each category. There are a total of 12 classes of products.

For each category of product images, various types of potential defects, denoted as T_c, exist. Here, $T_c = \{t_i | i \in [1, ..., D_c]\}$, and D_c is a function of the product class type c. For example, the product "candle" has $D_{c=1} = 7$ potential defects, including extra wax, foreign particles, missing wax chunks, unusual candle wicks, damaged packaging corners, different color spots, and wax melded out of the candle.

The benchmarked performance is reported using metrics including Image-AUROC, Image-F1score, Pixel-AUROC and Pixel-F1score.

B. Data collection strategy

High-quality data is often more pivotal than sheer volume in enhancing model performance, rendering the necessity of big data less absolute. In the context of the Visual Anomaly

detection (VisA) dataset, experiments are conducted to unveil redundancy within the dataset by varying the sizes of the training normal data. Specifically, the dataset size ranges from 10% to 100% of the original dataset size, with increments of 10%. The performance evaluation spans 12 products, assessed through four metrics: Image AUROC, Image F1-Score, Pixel AUROC, and Pixel F1-Score, and is visually represented in corresponding figures.

The performance figure for Padim with various dataset sizes is illustrated in Figure 2a. Instances where the performance exhibits an increase beyond 10% (comparing 10% vs 100% dataset size) are delineated with green border lines. Notably, only a few products, under certain metrics, demonstrate substantial performance enhancements. Particularly, metrics such as pixel_auroc and image_f1score show no significant improvements. The most notable performance boost is observed in the product "macaroni1," registering a 71.12% improvement in the pixel-f1score metric when transitioning from a 10% to the full dataset size.

A parallel experiment is conducted with Patchcore, as depicted in Figure 2b. However, the results are less promising, with only the "capsules" product showing a performance boost of 22.52% in the pixel_f1score metric. This outcome is primarily attributed to the coreset selection process of the algorithm, which inherently selects the most representative patch features during anomaly detection. The addition of data lacking additional variances does not lead to a noticeable improvement in performance. Consequently, this study advocates for a refined data collection strategy - Algorithm 1.

Algorithm 1: Data Collection Algorithm

Input: Initial dataset S, threshold T, required dataset size N
Output: Final dataset D with size N
Initialize: $D \leftarrow S$ Initialize dataset with golden set S
while $|D| < N$ **do**
  $newImages \leftarrow$ GENERATEIMAGES;
  $distances \leftarrow$ [COMPUTEDISTANCE$(newImage, s) \quad \forall s \in S, \forall newImage \in newImages$];
  $rankedIndices \leftarrow$ RANKDATA$(distances)$ Rank indices based on distances to S;
  $selectedImages \leftarrow$ $[newImages[i] \quad$ for i in rankedIndices$[25\% : 75\%]]$ Select images from 25th to 75th percentile;
  $D \leftarrow D \cup \{selectedImages\}$ Accept the selected images;
Return D

In particular, to measure the similarity of the generated images with the golden set, the feature extractor of ResNet18 is adopted to generate the feature embedding. To ensure the features maintaining the global information, the embedding from layer3 is utilized (refer to anomalib [46] for details). The similarity/distance is subsequently defined as the minimum of L1 distance from the generated image I^g to the golden set S,

$$d = \min_{s \in S} \sum_{0<i<N, 0<j<M, 0<c<C} \|I^g_{i,j,c} - s_{i,j,c}\|_1, \quad (1)$$

where N and M are the spatial dimension of features and C is the channel dimension.

In summary, the data collection strategy involves incorporating images that maintain a measured distance to the nearest sample in the golden set, striking a balance between not being too distant or too proximate. The calculation of similarity or distance is consistently applied to a fixed dataset size, ensuring that the time complexity of data collection remains constant for each sample.

IV. EXPERIMENTS

Experiments conducted in this study aim to validate the proposed data collection strategy, assess the resulted model performance improvements, and the impact of data diversification through diffusion models. For the sake of maintaining generality, the model utilized in this study is modified *Padim* architecture, by enlarging the feature dimensions so that the added variance by diffusion models are not randomly dropped. The performances are benchmarked on the VisA dataset.

A. Dataset Description

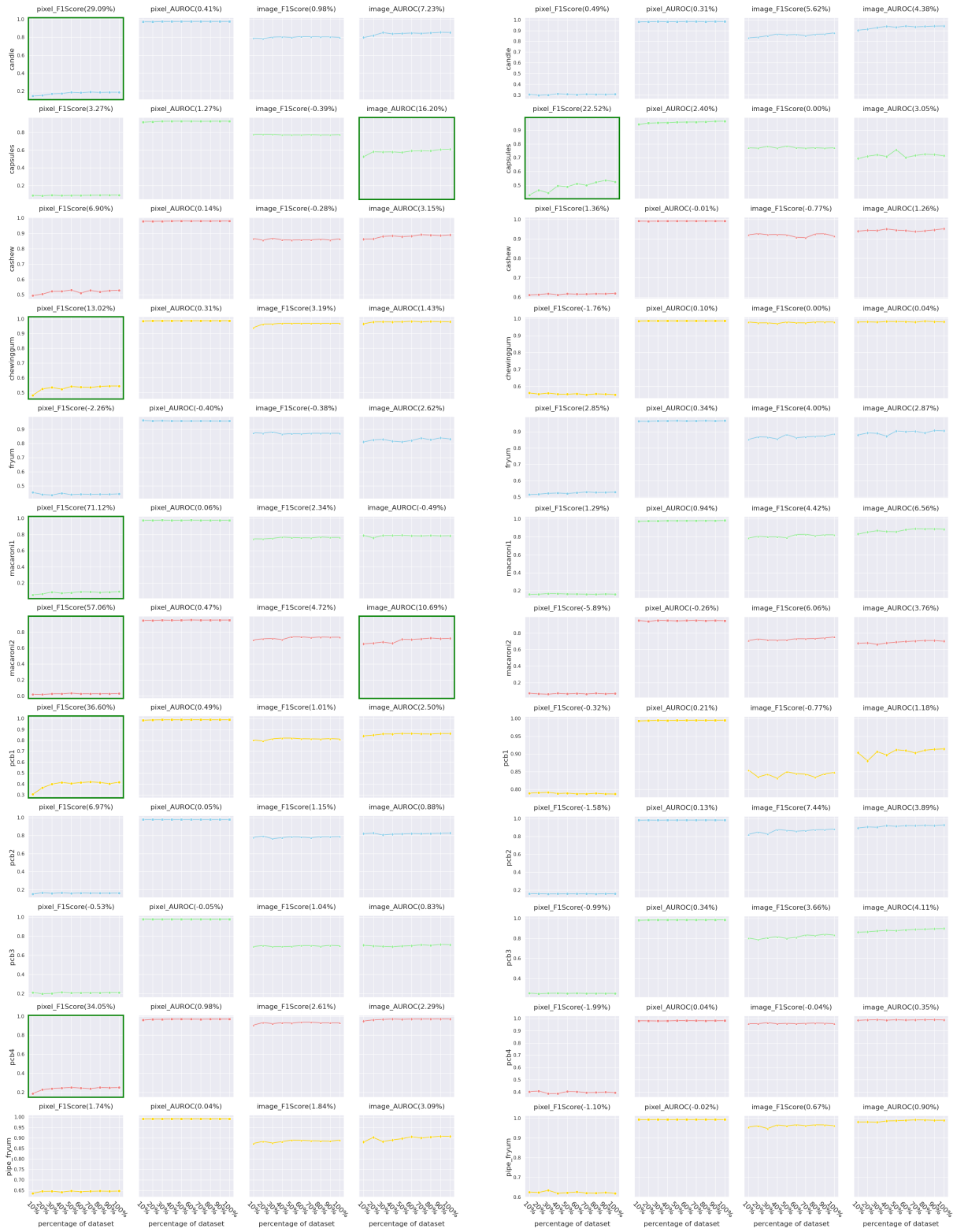
The VisA dataset comprises 10,821 high-resolution color images, including 9,621 normal and 1,200 anomalous samples, encompassing 12 objects across 3 domains [47]. The dataset covers objects with complex structures (PCB1, PCB2, PCB3, PCB4), images with multiple instances (Capsules, Candle, Macaroni1, Macaroni2), and images with single instances (Cashew, Chewing gum, Fryum, Pipe fryum). Anomaly types encompass surface defects like scratches, dents, color spots, or cracks, as well as structural defects such as misplacements or missing parts. Additionally, some images can contain multiple defects.

B. Unsupervised modelling

This study centers on examining the influence of the dataset on the performance of modeling in an unsupervised setting. The data-centric approach is demonstrated to be effective, irrespective of the modeling approaches employed. Padim [6] utilizes a pretrained Convolutional Neural Network (CNN) to extract patch embeddings and employs the multivariate Gaussian distribution to obtain a probabilistic representation of the normal class. Given a set of N normal images I, the patch embedding vectors located at position (i, j) are defined as $X_{ij} = \{x^k_{i,j}, k \in [1, N]\}$. Subsequently, the two hyperparameters μ_{ij} and Σ_{ij} are calculated by

$$\mu_{ij} = \frac{1}{N} \sum_k x^k_{ij}, \ where \ k \in [1, N] \quad (2)$$

$$\Sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{x^k_{ij}} - \mu_{ij})(\mathbf{x^k_{ij}} - \mu_{ij})^T + \epsilon I \quad (3)$$



(a) Padim performances on VisA dataset with various sizes of dataset

(b) PatchCore performances on VisA dataset with various sizes of dataset

Fig. 2: Performance comparison on VisA dataset with various sizes of dataset

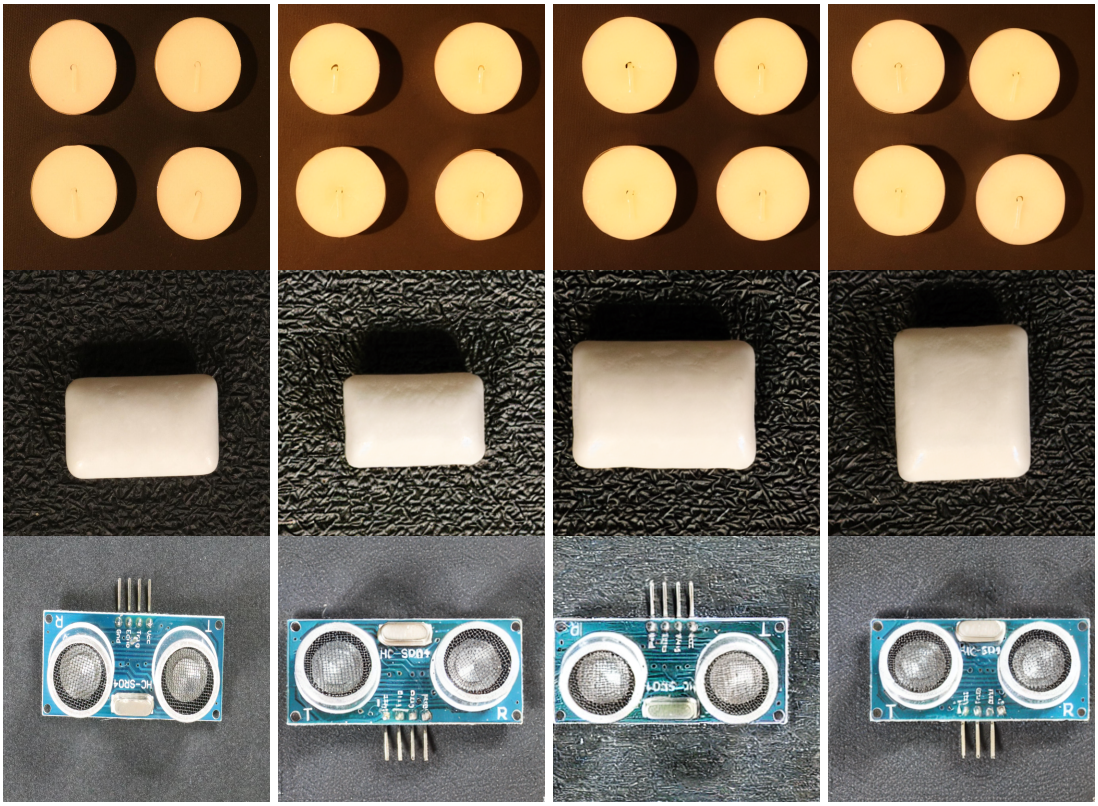


Fig. 3: Diffusion generated images

With the pre-calculated multivariate Gaussian Distribution $\mathcal{N}(\mu_{ij}, \mathbf{\Sigma_{ij}})$, given a test image I_t, the anomaly score at pixel (i, j) is derived as the Mahalanobis distance

$$\mathcal{M}(x_{ij}) = \sqrt{(x_{ij} - \mu_{ij})^T \Sigma_{ij}^{-1} (x_{ij} - \mu_{ij})} \tag{4}$$

The effectiveness of anomaly detection relies heavily on the representativeness of the constructed multivariate Gaussian distribution. In order to enhance the diversity of the dataset while maintaining control over how much the generated dataset deviates from the original, image generation using a fine-tuned diffusion model is introduced.

C. Data diversification with fine tuned diffusion models

To empower the diffusion model to generate domain-specific images, such as candles, capsules, PCB boards, etc., and introduce **controlled** variance into the dataset, the prior preservation loss term is omitted. Instead, only the reconstruction loss conditioned on the textual embedding is retained.

$$E_{\mathbf{x},\mathbf{c},\epsilon,\epsilon',t,\Omega}([\omega_t \|\hat{\mathbf{x}}_\theta(\alpha_\mathbf{t}\mathbf{x} + \delta_\mathbf{t}\epsilon, \mathbf{c}) - \mathbf{x}\|_2^2]),\ \Omega \in \mathbf{FG}(I) \tag{5}$$

In this context, $\mathbf{c}$ represents the product name c formatted as "a photo of c". This ensures that the generated images contribute controlled variance to the dataset without significantly deviating from the original distribution. Additionally, a foreground-aware term denoted as the extra constraint term Ω is introduced, where **FG** is the foreground extraction algorithm based on GroundingDINO [48] and SAM [49]. The foreground bounding box is firstly predicted with GroundingDINO prompted with the product name. The bounding box is then sent to SAM for segmentation. Foreground-aware data enrichment with image diffusion processes has been demonstrated to enhance model performance while imposing fewer constraints on the dataset size.

Samples with diffusion-generated normal images are showcased in Fig. 3, featuring sample images from each of the three domains.

For each of the three domains, images equivalent to 10% of the original dataset size are generated. For example, for candles, approximately 90 images are generated to represent 10% of the total normal images within this category. In summary, the experiment compares the performance with generated images to scenarios with 40% of the original dataset and 100% of the original dataset. The results of the proposed algorithm DCADDM, padim and patchcore are summarized in Table. I.

In the context of multiple-instance scenarios, the proposed algorithm exhibits a significant enhancement in anomaly localization performance. Notably, the pixel-level Image F1-Score and Image AUROC achieve values of 0.9458 and 0.9833 respectively, surpassing *Patchcore* (employing the entire dataset) by approximately 0.07 and 0.04.

In the realm of single-instance scenarios, the proposed algorithm surpassed all alternative methods based on Image AUROC (0.9906) and Pixel F1-Score (0.6001). It attains comparable performance with a reduced dataset, as evidenced by Pixel AUROC (0.9700) and Image F1-Score (0.9703).

In scenarios involving products with complex structures or textures, the proposed algorithm outperforms both *Padim* and

Domain	Product	Setting	Pixel F1-Score	Pixel AUROC	Image F1-Score	Image AUROC
multiple instances	*candle*	*padim + 100% dataset*	0.1873	0.9756	0.7982	0.8532
		patchcore + 100% dataset	0.3086	**0.9837**	0.8785	0.9423
		padim + 40% dataset	0.1728	0.9745	0.8037	0.8380
		patchcore + 40% dataset	**0.3122**	0.9818	0.8673	0.9389
		ours + 30% dataset + 10% diffusion	0.2866	0.9742	**0.9458**	**0.9833**
single instance	*chewinggum*	*padim + 100% dataset*	0.5446	**0.9875**	0.9700	0.9802
		patchcore + 100% dataset	0.5508	0.9859	**0.9796**	0.9802
		padim + 40% dataset	0.5233	0.9870	0.9700	0.9790
		patchcore + 40% dataset	0.5540	0.9862	0.9700	0.9826
		ours + 30% dataset + 10% diffusion	**0.6001**	0.9700	0.9703	**0.9906**
complex structure	*pcb1*	*padim + 100% dataset*	0.4168	0.9883	0.8117	0.8615
		patchcore + 100% dataset	0.7870	**0.9951**	**0.8479**	**0.9141**
		padim + 40% dataset	0.4139	0.9881	0.8201	0.8576
		patchcore + 40% dataset	**0.7884**	0.9941	0.8317	0.8970
		ours + 30% dataset + 10% diffusion	0.5763	0.9482	0.8468	0.9070

TABLE I: Benchmarking Proposed DCADDM with Padim and Patchcore on VisA

Patchcore when trained with an equivalent dataset size (40% of the original dataset), demonstrated through higher Image-level F1-Score (0.8468 vs. 0.8317) and AUROC (0.9070 vs. 0.8970) metrics. When it comes to the anomaly localization capability measured by Pixel F1-Score, *Patchcore + 40% dataset* outperforms the proposed solution (0.7884 vs. 0.5763). This is due to the challenges in generating high-quality product images in complex texture scenario.

V. Ablation Studies

In the ablation study, the influence of the foreground-aware factor on model performance is scrutinized, as depicted in Fig. 4. The assessment encompassed three diverse products spanning various domains, measuring Image-level and Pixel-level F1-Score, as well as AUROC, under both diffusion and diffusion + fg configurations.

Notably, the "candle" product demonstrated a substantial performance enhancement when the foreground-aware factor was introduced, particularly evident in the metrics of Image-level AUROC and F1-Score. This improvement is attributed to the effective extraction of the foreground, ensuring that variability is introduced solely to the foreground, minimizing background interference.

Conversely, Pcb1 experienced a marginal performance dip with the addition of the foreground-aware factor. The intricate texture within Pcb1 product images heightened the challenge of accurate foreground extraction, consequently introducing noise to the anomaly detection results.

In the case of the "chewinggum" product, no significant difference was observed. The distinct boundary between foreground and background allowed the model to naturally prioritize the foreground, even in the absence of the foreground-aware factor.

In summary, the foreground-aware factor proves most effective in a multiple-instance scenario, yielding notable performance improvements. However, in scenarios involving products with complex structures, it may lead to a performance drop under certain metrics. Nevertheless, the incorporation of a foreground-aware factor in the diffusion model for data augmentation remains advantageous.

VI. Conclusion

In conclusion, the exploration into anomaly detection, a pivotal aspect in identifying deviations within product images, has culminated in the development of Data-centric Anomaly Detection with Diffusion Models. The proposed algorithm encompasses a comprehensive strategy for data collection in unsupervised modeling settings. Additionally, the demonstrated effectiveness of data diversification with fine-tuned diffusion models has proven instrumental in enhancing model performance. This enhancement stems from the introduced variance into the training dataset, leveraging the generative capabilities of AI. Furthermore, the foreground-aware factor incorporated into the diffusion process imposes constraints on the location of introduced variance, a component whose significance is underscored in the ablation studies. Through the experiments conducted in this paper, it is concluded that, in anomaly detection with unsupervised modeling approaches, the necessity for big data is not absolute. Instead, the emphasis lies on high-quality data with substantial diversity, which is crucial for achieving satisfactory performances. Text-to-image generation with diffusion models emerges as a qualified candidate for generating a diversified dataset.

Appendix A

More sample images comparing the proposed algorithm vs. *Padim* and *Patchcore* shown in Fig. 5.

References

[1] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 1

[2] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, 14(9):3098–3104, 2017. 1

[3] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011. 1

[4] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1519, 2015. 1

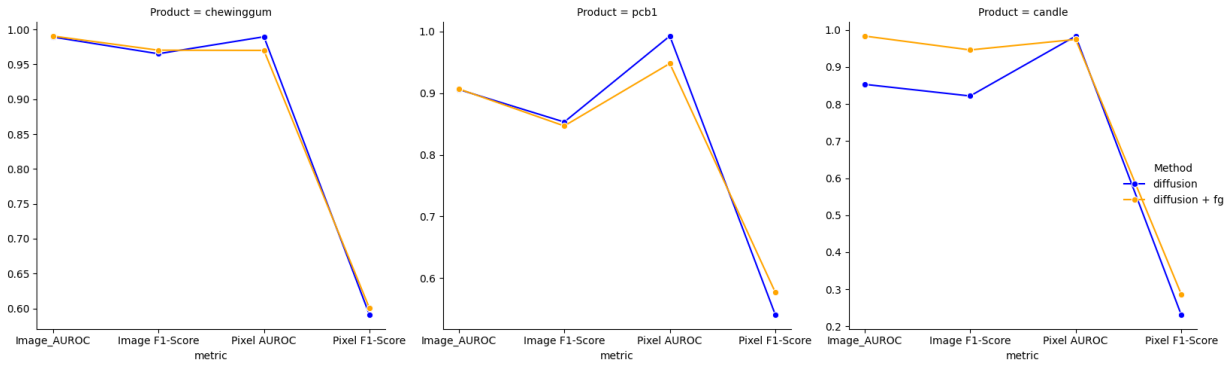


Fig. 4: Impact of fg factor

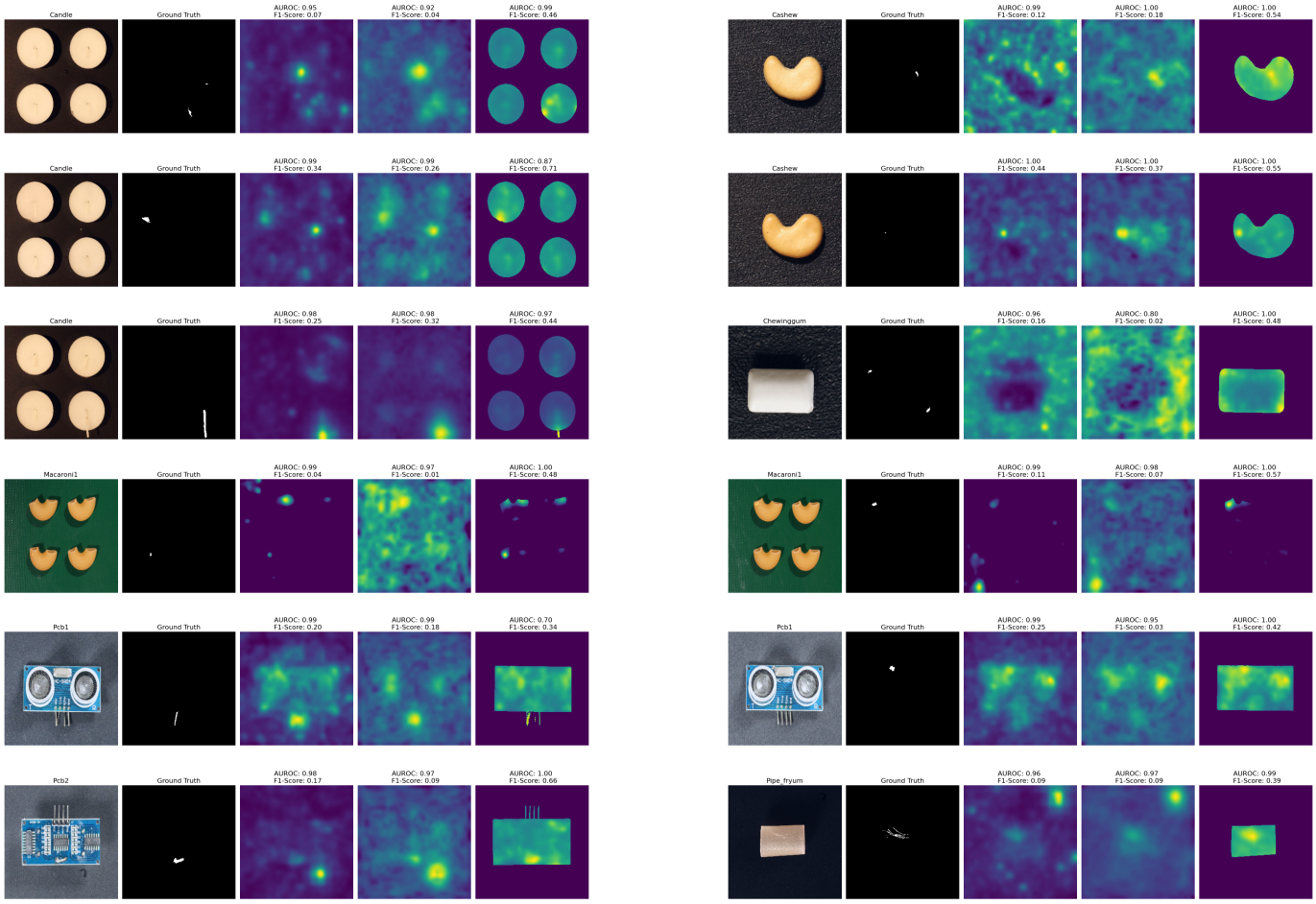


Fig. 5: DCADDM vs. Padim & Patchcore

[5] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. *CoRR*, abs/2106.08265, 2021. 1, 2

[6] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. *CoRR*, abs/2011.08785, 2020. 1, 2, 3

[7] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, pages 4800–4809. PMLR, 2019. 1

[8] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pages 4–11, 2014. 1

[9] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019. 1

[10] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018. 1

[11] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *arXiv preprint arXiv:2303.14535*, 2023. 1

[12] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[14] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. 1

[15] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 1–17. Springer, 2020. 1

[16] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 1

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1, 2

[18] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2

[19] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[21] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018. 2

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

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[24] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2

[25] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 2

[26] Anne-Sophie Collin and Christophe De Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7915–7922. IEEE, 2021. 2

[27] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2

[28] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer, 2020. 2

[29] Thomas Schlegl, Philipp Seebock, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurthb. Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2, 2017. 2

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[32] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2

[33] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022. 2

[34] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 2

[35] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022. 2

[36] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[37] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. 2021. 2

[38] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[39] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[40] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[43] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 2

[44] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[46] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection, 2022. 3

[47] 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*, pages 392–408. Springer, 2022. 3

[48] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[49] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5