HiClassGen: High-Resolution Image Augmentation with Class and Shape Controllable Diffusion Models

Adnan Md Tayeb, Jae-Min Lee, Dong-Seong Kim
Networked Systems Laboratory, Department of IT Convergence Engineering,
Kumoh National Institute of Technology, Gumi, South Korea.
(mdtayebadnan, ljmpaul, dskim)@kumoh.ac.kr

***Abstract*—Deep learning models generally require large quantities of data to ensure effective generalization and capture complex patterns in diverse scenarios. However, acquiring sufficient labeled data is often costly and time-consuming in many domains, such as medical imaging, industrial defects, remote sensing, and other specialized fields. To address this challenge, we present HiClassGen, a novel image augmentation framework that leverages class- and shape-controllable diffusion models to generate high-quality, domain-specific synthetic data. HiClassGen integrates class-based semantic guidance and shape constraints during the image generation process, enabling precise control over both the augmented images' class labels and morphological features. This customization ensures that the synthetic data generated aligns with the specific needs of the target task, leading to better performance and generalization. Additionally, HiClassGen automatically generates corresponding annotations for the synthetic data, ensuring seamless integration into supervised learning pipelines. Experiments demonstrate the effectiveness of our proposed model, showing a 10% increase in accuracy on benchmark datasets.**

***Index Terms*—Generative AI, Controllable Diffusion, Data Augmentation,**

I. Introduction

Deep learning models have demonstrated remarkable performance across various domains, ranging from computer vision and natural language processing to medical imaging and autonomous systems. However, a major challenge in applying deep learning techniques effectively is the need for large, high-quality datasets to ensure good generalization. Models with higher capacities can learn complex patterns and representations, but without sufficient data, they are prone to overfitting, resulting in poor performance on unseen data. In many real-world applications, especially in specialized fields like medical imaging, remote sensing, and scientific research, collecting large labeled datasets is often impractical, expensive, and time-consuming. This data scarcity presents a significant barrier to fully leveraging the potential of deep learning models [14], [24].

Data augmentation has emerged as a powerful technique to address the challenge of annotating new datasets by generating additional training examples from existing data. This method is widely used in object detection and segmentation tasks, where it helps improve the model's ability to generalize by introducing variability in the input data. Traditional augmentation techniques, such as rotation, scaling, flipping, and other image manipulations, allow the model to focus on learning more invariant features, ultimately enhancing the robustness of the trained model. In recent years, more advanced augmentation strategies have been introduced, including image-erasing methods such as random erasing [30], GridMask [2], FenceMask [11], and Cutout [4], which selectively obscure parts of the image to force the model to learn features that are not reliant on specific regions. Image mix methods, such as Mosaic [5], Mixup [28], and CutMix [27], combine multiple images or portions of images to create more diverse training examples, further improving the model's ability to handle complex real-world scenarios. Copy-paste techniques, where image samples are replicated [6] and integrated into new contexts, are also gaining traction as they offer effective ways to augment datasets with minimal computational cost. Moreover, the latest advancements in generative models, such as CLIP and Stable Diffusion, have revolutionized data augmentation by enabling the creation of entirely synthetic images that are indistinguishable from real ones. These generative methods not only provide an innovative way to expand datasets but also contribute to creating diverse, high-quality examples that push the boundaries of traditional data augmentation techniques, further enhancing the performance and adaptability of models in challenging tasks.

However, generative data augmentation presents unique challenges when dealing with bounding box annotations or semantic masks [22], [23], unlike traditional augmentation techniques where bounding boxes or semantic masks can be easily manipulated through simple transformations. This complexity has confined generative methods mainly to image classification tasks, where bounding boxes or semantic masks are not integral. However, integrating generative augmentation becomes more difficult in object detection and semantic segmentation, where precise annotation labels are crucial. Although some specialized generative models aim to handle this issue, they have notable limitations. One example is layout-to-image models, such as those found in [3], [10], [29], which require dense bounding box distributions for training. These models are unsuitable for object detection tasks as they fail to provide flexibility. While [1] has explored the potential of layout-to-image models for data augmentation in object detection datasets, their experiments remain difficult to replicate due to code and dataset access restrictions, limiting further exploration in this area. Another approach is using copy-paste methods with diffusion models, where the target objects are

generated based on textual descriptions and then extracted using external segmentation models [13], [17], [18], [21] before being pasted onto real images. However, this approach often results in less realistic generated data, as the objects lack the fine details of natural images, impacting their usefulness in practical applications. Moreover, both strategies necessitate training on object detection or segmentation datasets, which can be resource-intensive and may not transfer well to different domains. These challenges underscore the ongoing need for innovative solutions to make generative augmentation viable for object detection and segmentation tasks while simultaneously improving the realism and flexibility of the generated data.

In this work, we propose HiClassGen, a novel image augmentation framework that combines the power of class- and shape-controllable diffusion models to generate high-resolution, domain-specific synthetic images. Unlike traditional methods, HiClassGen allows for fine-grained control over the class labels and morphological features of the generated images, ensuring that the synthetic data aligns closely with the needs of downstream tasks. We explore the benefits of this approach by evaluating its performance on multiple datasets, demonstrating improvements in segmentation tasks. By enabling precise control over the data generation process, HiClassGen provides a powerful tool for overcoming the data scarcity problem, offering a scalable and effective solution for enhancing the performance of deep learning models in data-limited domains.

To summarize, our contributions are as follows:

- We introduce a novel model, HiClassGen, designed to generate annotated images, providing an effective solution for robust data augmentation.
- We integrate ShapeNet into the model to precisely control the shape of generated objects, ensuring the accuracy of annotations.
- We incorporate a class control mechanism that allows the model to generate class-specific images, thereby enhancing its versatility in handling diverse object categories.

II. Preliminary

A. Diffusion Models

Diffusion models (DDMs) [8], [9], [15], [16], [20], [26] are probabilistic models designed to generate data by progressively denoising normally distributed variables. Training these models involves a forward diffusion process that adds Gaussian noise to an image x_0 over a series of steps, forming a Markov Chain of length T. The objective is to train the model to reverse this diffusion process, reconstructing clean samples from pure noise through the reverse diffusion process.

Given a sample x_0 drawn from the real data distribution $q(x)$, the forward diffusion process incrementally adds Gaussian noise to x_0 at each step, governed by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. This can be formalized as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \tag{1}$$

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \tag{2}$$

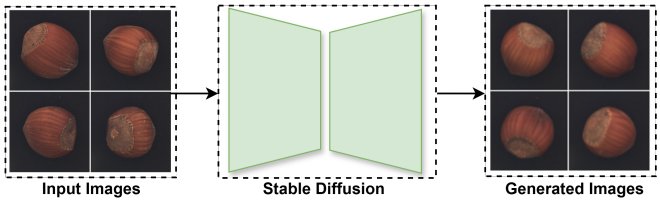

Fig. 1. Structure of the stable diffusion to generate anomaly-free images

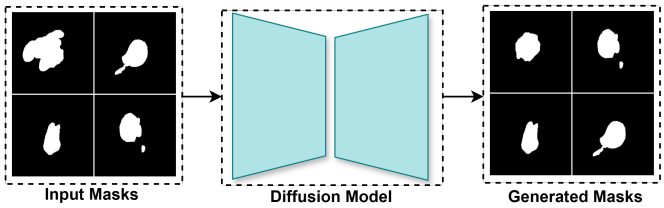

Fig. 2. Structure of the diffusion model to generate random masks

For sufficiently large T, x_T approximates an isotropic Gaussian distribution.

In the reverse diffusion process, the model aims to undo the noising steps, mapping noisy samples back to clean images. This is formalized as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \tag{3}$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{4}$$

For simplicity, practitioners often assume $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$.

The model's output is frequently structured as a residual, predicting the noise ϵ_t added at each step rather than directly generating the image. The reconstructed image x_0 is obtained by subtracting ϵ_t from x_t. Thus, the training objective is:

$$L_t = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2\right] \tag{5}$$

where L_t represents the loss function at time step t.

B. Stable Diffusion

Latent Diffusion Models (LDMs) [19] represent an advanced variant of diffusion models (DDMs) that operate within a compressed latent space rather than the high-dimensional pixel space. This approach significantly reduces the computational demands of both training and inference. In LDMs, an encoder $E(\cdot)$ transforms the input image into a compact latent representation z, where diffusion and denoising processes occur. The decoder $D(\cdot)$ subsequently reconstructs the image from this latent representation, yielding $\tilde{x} = D(z)$.

Stable Diffusion is a notable instance of an LDM, leveraging cross-attention mechanisms to accommodate diverse conditioning signals, such as textual inputs. Its training framework incorporates two essential regularization strategies: (i) KL regularization, which aligns the latent space distribution with a Standard Normal distribution, and (ii) VQ regularization, which integrates a vector quantization layer into the decoder

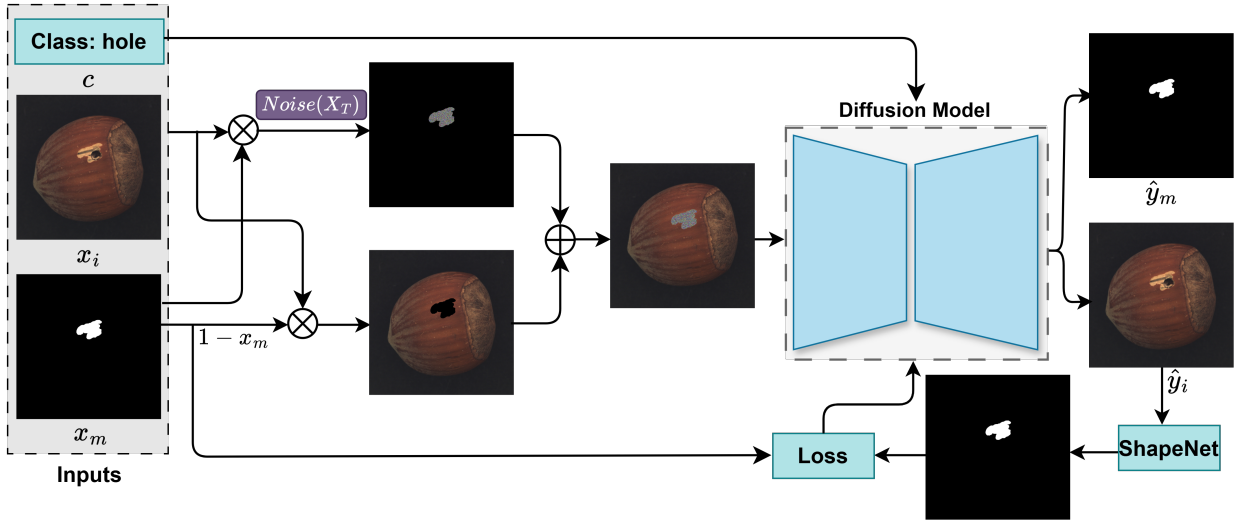

Fig. 3. Overview of the proposed HiClassGen

architecture (similar to VQGAN but with quantization embedded in the decoder). Pre-trained on an extensive dataset of natural images, Stable Diffusion has demonstrated exceptional performance across a variety of tasks. Moreover, its publicly available pre-trained weights make it a versatile base model for numerous downstream applications.

III. Method

Given a class condition and c an input image $x_i \in \mathbb{R}^{W \times H \times C}$, a binary mask $x_m \in \{0,1\}^{W \times H}$ where a value of 1 identifies the foreground regions and 0 indicates background. The main goal is to generate new images $\hat{y} \in \mathbb{R}^{W \times H \times C}$ along with the corresponding annotations $\hat{y}_m \in \{0,1\}^{W \times H}$. We followed SmartBrush [25] and utilized it as the base of our method.

Algorithm 1 Training of HiClassGen

1: **Input:** Input image x_i, corresponding mask x_m, Class c
2: **Output:** Image x_o, Mask x_m
3: Train Stable Diffusion with good Images:
  Sample $z_0 \sim E(x_i)$
  $z_T \sim \text{noise}(z_0, T)$: $x_t = \sqrt{1-\beta_t} \cdot x_{t-1} + \sqrt{\beta_t} \cdot \epsilon$
  for all t from T to 0 do
    Denoise (z_T, t): $x_{t-1} = x_t - \beta_t \cdot \epsilon_\theta(x_t, t)$
  end for
  $x_q = D(z_0)$
4: Training HiClassGen with ShapeNet
  $\tilde{x}_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
  $x_t = \tilde{x}_t \odot m + x_0 \odot (1-m)$
  $\mathcal{L}_{\text{DM}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(x_t, t, x_m, c)\|_2^2 \right]$. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DM}} + \lambda \mathcal{L}_{\text{ShapeNet}}$
5: **Return** x_o, x_m

A. Training on good Images

In the initial phase of our methodology, we focus on utilizing a stable diffusion model as illustrated in Figure 1. This phase aims to harness the advanced generative capabilities of stable diffusion to produce a diverse set of good images, forming the foundation for the subsequent generation of class-specific images. Stable Diffusion has consistently demonstrated exceptional proficiency in generating high-quality, photorealistic images across various domains. By incorporating it into HiClassGen, we ensure that the generated images exhibit high fidelity and detailed visual consistency, aligning closely with real-world standards.

B. Generate random mask

To generate images along with their annotated masks, HiClassGen requires input sample masks corresponding to their images. During the training phase, existing annotated masks are utilized to learn the relationship between masks and image features. However, during the image generation phase, we input a random mask to create new synthetic data.

To produce these random masks, we train a diffusion model exclusively on the existing annotated masks. This ensures that the generated masks retain realistic structural and morphological characteristics specific to the target domain, while also introducing variability necessary for effective data augmentation. The diffusion model leverages its ability to learn high-dimensional patterns in the annotated mask data to generate plausible and structurally coherent masks.

As illustrated in Figure 2, the diffusion model generates random masks that are diverse yet exhibit realistic features, maintaining morphological similarity to the input masks used during training. This balance of diversity and realism ensures that the generated masks are not only domain-specific but also representative of potential variations within the dataset. By inputting these generated random masks into the augmentation pipeline, HiClassGen creates high-quality synthetic images with precise annotations, significantly enriching the dataset for downstream tasks.

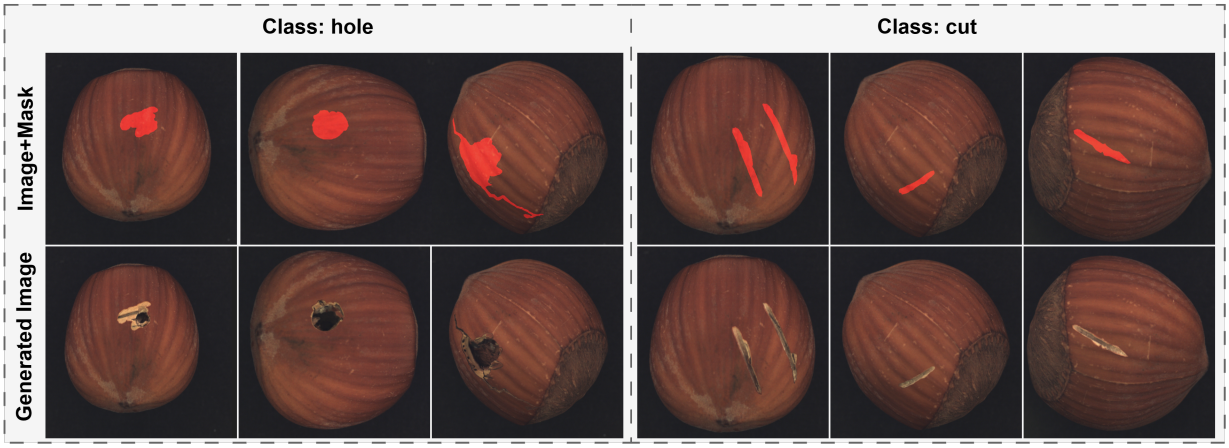

Fig. 4. Examples of the generated image leveraging the image, mask, and class conditions

C. Class and shape-controlled Generation

We have utilized the annotated segmentation dataset to train the model for this phase. These datasets consist of annotated masks $\{m_i\}_{i=1}^{N}$, where N represents the total number of annotations, and each masked region $x \odot m_i$ contains a single object. Each mask m_i is associated with a corresponding class label c_i, such as "hole" or "wrinkle."

During the forward process, a segmentation mask m and its corresponding class text label c are randomly selected for a given image x. We initialize $x_0 = x$ and introduce noise only within the masked region, leaving the remaining pixels untouched. The noisy image is defined as:

$$\tilde{x}_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon,$$

$$x_t = \tilde{x}_t \odot m + x_0 \odot (1 - m),$$

where $\epsilon \sim \mathcal{N}(0, I)$, t represents the timestep in the forward process, and $\bar{\alpha}_t$ controls the noise scale.

The model takes x_t, m, and c as inputs, enabling it to leverage the clean background information while learning to process and reconstruct the noisy masked region effectively. To recover the masked region $x_0 \odot m$, we ensure that the generated objects in the foreground m are consistent with the background. We train a network ϵ_θ to predict the noise ϵ from the noisy input x_t:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(x_t, t, m, c)\|_2^2 \right].$$

In the inference phase, we initialize the masked region with random Gaussian noise as:

$$x_T = \epsilon \odot m + x_0 \odot (1 - m),$$

where T is the number of sampling steps. The diffusion process is then reversed to obtain the inpainted result x_0.

D. Shape Control

Our training masks are derived from segmentation annotations, ensuring that they are accurate instance masks. Training the model with these precise masks encourages it to faithfully adhere to the shape of the input mask during testing. To further refine shape precision, our method incorporates a segmentation model, which we call ShapeNet, as illustrated in Figure 3.

First, we train the ShapeNet using the existing annotated images to learn the shapes of specific classes. During the training of HiClassGen, ShapeNet predicts the shapes of the generated images and compares them with the ground truth masks to compute the loss values. The computed loss is then passed to the Stable Diffusion model, scaled by a factor α, to guide its learning process. This integration enables the model to learn the shape more accurately.

Through this approach, we can control whether the generated objects align with the input masks by specifying different mask precision indicators.

IV. Experiment

To evaluate the effectiveness of HiClassGen, we conducted experiments using the MVTec AD dataset, encompassing both the defect image generation task and the downstream defect recognition task.

A. Dataset: MVTec AD

The MVTec Anomaly Detection dataset (MVTec AD) is an openly available benchmark dataset containing ten object categories and five texture categories commonly observed in industrial applications. Each object/texture category includes up to eight defect types, with pixel-level masks marking the defective regions. Originally designed for defect localization, MVTec AD is well-suited for few-shot defect image generation, as most object/texture categories include 200–400 defect-free samples and 10–25 defect images per defect category.

In the first stage, the backbone of HiClassGen is trained using defect-free images from a specific object or texture category in the training set. This step focuses on augmenting the

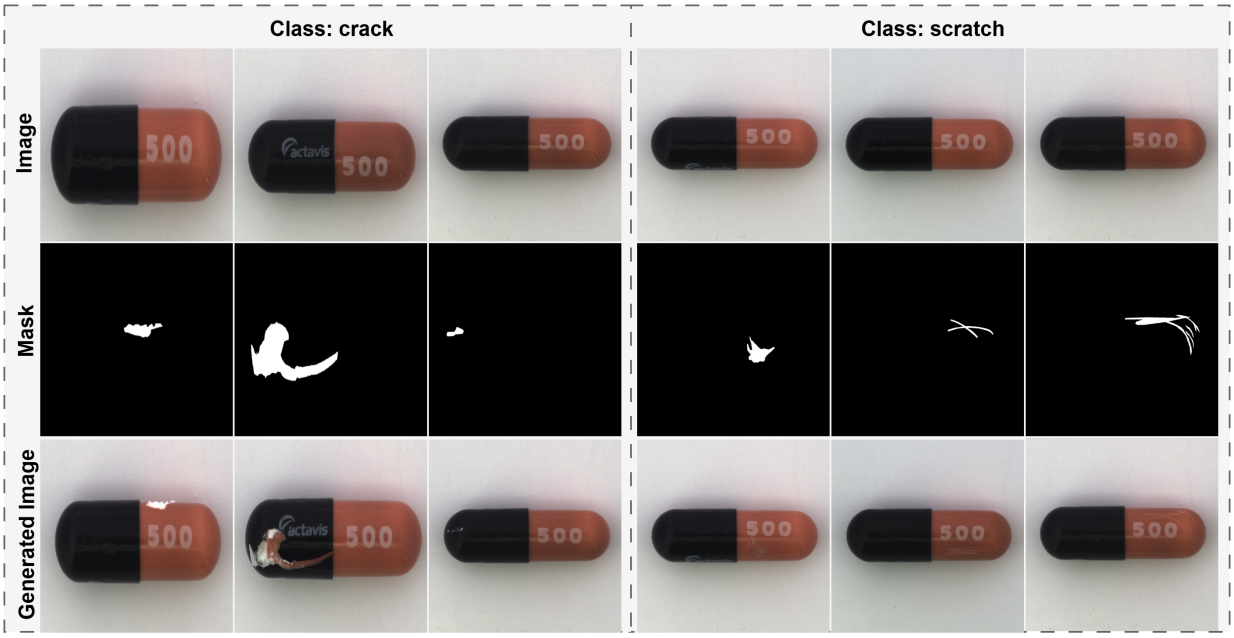

Fig. 5. Examples of the generated image leveraging the image, mask, and class conditions

number of defect-free images to provide a robust foundation for the model. To ensure the model learns diverse variations, defect images are first enhanced through a blending process before being used for training. This approach allows HiClassGen to effectively capture the nuances of defect patterns while maintaining the integrity of defect-free samples.

In the second stage, a distinct HiClassGen model is trained for each defect category associated with that object or texture category. All images are resized to a 225×225 resolution to ensure computational efficiency.

In the main paper, we primarily focus on the object category *hazelnut*, which poses a significant challenge due to its naturally high variability and complex appearance compared to other manufactured objects. This category contains four defect types: crack, cut, hole, and print. Results for other categories are included in the supplementary material.

B. Implementation Details

Our network was implemented using Python 3.8.10 and PyTorch 1.13.1, and experiments were performed on a system equipped with four NVIDIA GeForce RTX 3090 GPUs, each providing 24GB of memory. For parameter optimization, we utilized the AdamW optimizer [12] with a learning rate of 1×10^{-5}, along with beta values of 0.9 and 0.999. The input resolution was configured at 225×225, and the training process employed a batch size of 16.

C. Result Analysis

Our main objective in this study is to generate high quality synthetic images for use in industries. Since annotating data is costly, the generated images must contain the annotations of the defect's location exactly. According to Figure 4 and 5, our approach successfully generates realistic images. The generated images' appearance closely matches the input guide's geometries. Furthermore, the defects in these images match those shown in the provided masks.

TABLE I
AVERAGE RESULTS FOR DEFECT DETECTION ON STEEL SURFACES, SPECIFICALLY FOR WRINKLE AND NOZZLE DEFECTS.

Used data	mAP	Precision	Recall
Real only	45.54	48.12	42.71
Synthetic only	32.32	40.25	35.7
Real + Synthetic	55.45	63.89	52.35

To assess the quality of the generated data, we employed the Frechet Inception Distance (FID) metric [7]. When comparing the distribution of the synthetic images to that of the Steel surface dataset, an FID score of 99.57 was achieved. Additionally, we explored the potential application of these annotated synthetic images in an industrial context by leveraging the widely-used object detection model, YoLo-NAS. The YoLo-NAS model was trained under four distinct scenarios: (i) using only real data for training; (ii) using real data augmented with basic augmentation techniques; (iii) training exclusively on synthetic data; and (iv) pre-training on synthetic data followed by fine-tuning with real data. Evaluation on a separate test dataset demonstrated that incorporating synthetic data consistently enhanced defect detection across all evaluated metrics. The quantitative results, as presented in I, reveal that combining real and synthetic data significantly outperformed training with either type of data alone. Notably, the mean average precision (mAP) increased by approximately 10% when both real and synthetic datasets were utilized, underscoring the effectiveness of synthetic images in boosting model performance.

V. Conclusion

In this work, we proposed HiClassGen, a novel framework for generating high-quality annotated synthetic images to address challenges in data augmentation and improve model performance in industrial defect detection tasks. By integrating ShapeNet for precise shape control and employing a class control mechanism, HiClassGen enables the generation of diverse, realistic, and class-specific synthetic images. Our experimental results demonstrated that combining real and synthetic data significantly enhances defect detection performance, with a notable improvement in mean average precision (mAP) compared to using only real or synthetic data. Furthermore, the Frechet Inception Distance (FID) score validated the authenticity of the generated images. These findings underscore the potential of synthetic data to overcome data scarcity and enhance model robustness in real-world industrial applications. Future work will focus on extending HiClassGen to other domains, exploring additional generative techniques, and deploying the framework in practical industrial settings for comprehensive evaluation.

Acknowledgment

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of IITP grant funded by the Korea government(MSIT) (IITP-2025-RS-2020-II201612, 33%) and by Priority Research Centers Program through the NRF funded by the MEST(2018R1A6A1A03024003, 33%) and by the MSIT, Korea, under the ITRC support program(IITP-2024-RS-2024-00438430, 34%)

References

[1] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arxiv:2306.04607*, 2023.

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.

[3] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.

[4] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[10] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023.

[11] Pu Li, Xiangyang Li, and Xiang Long. Fencemask: A data augmentation approach for pre-extracted image features. *arXiv preprint arXiv:2006.07877*, 2020.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[13] Timo Luddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096. IEEE, 2022.

[14] Adnan Md Tayeb and Tae-Hyong Kim. Unestformer: Enhancing decoders and skip connections with nested transformers for medical image segmentation. *IEEE Access*, 12:190996–191009, 2024.

[15] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[16] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[17] Nikita Pavlichenko and Dmitry Ustalov. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2067–2071, 2023.

[18] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2303–2316, 2022.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[20] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[21] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection, and video salient object detection. *IEEE Transactions on Multimedia*, 2023.

[22] A. M. Tayeb, K.-H. Lee, J.-Y. Seo, G.-W. Kim, H.-B. Park, H.-J. Lee, and T.-H. Kim. Facial skin condition analysis based on pore and wrinkle segmentation. In *Proceedings of the Symposium of the Korean Institute of Communications and Information Sciences*, pages 1648–1649, 2023.

[23] Adnan Md Tayeb, Mst Ayesha Khatun, Mohtasin Golam, Md Facklasur Rahaman, Ali Aouto, Oroceo Paul Angelo, Minseon Lee, Dong-Seong Kim, Jae-Min Lee, and Jung-Hyeon Kim. Smartrsd: An intelligent multimodal approach to real-time road surface detection for safe driving, 2024.

[24] Adnan Md Tayeb, Hope Leticia Nakayiza, Heejae Shin, Seungmin Lee, Chaesoo Lee, YeongHun Lee, Dong-Seong Kim, and Jae-Min Lee. DefectGen: Few-Shot Defect Image Generation Using Stable Diffusion for Steel Surface Analysis. In *2024 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 2092–2097. IEEE, 2024.

[25] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model, 2022.

[26] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[28] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[29] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.

[30] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008. AAAI Press, 2020.