Improving Semantic Segmentation Models through Synthetic Data Generation via Diffusion Models

Jonas Rabensteiner RABJONAS@GMAIL.COM
Free University of Bozen-Bolzano
Bolzano, BZ 39100, ITA

Cynthia I. Ugwu* CUGWU@UNIBZ.IT
Free University of Bozen-Bolzano
Bolzano, BZ 39100, ITA

Oswald Lanz OSWALD.LANZ@UNIBZ.IT
Free University of Bozen-Bolzano
Bolzano, BZ 39100, ITA

Reviewed on OpenReview: /

Editor: /

Abstract

It is often difficult to obtain industrial data for semantic segmentation due to the costs and time required for annotation. However, deep learning models perform poorly when trained on small datasets. The current advances in generative models can be exploited to enhance existing datasets with synthetic data. In semantic segmentation, generating images is not sufficient, as the images need to fit with corresponding labels on a pixel level, which makes data generation more challenging. Our work exploits Diffusion Models, to generate synthetic data for industrial semantic segmentation. Our thorough experimentation reveals that the generated data can contribute to improve the performance of semantic segmentation models without altering their architecture when trained on a mix of real and synthetic data. Additionally, some experiments demonstrate the feasibility of achieving improvements by exclusively training the models on synthetic data. The source code is available at `https://github.com/RabJon/isdm`.

Keywords: semantic image synthesis, semantic segmentation, diffusion models, anomaly localization, synthetic data generation

1 Introduction

Semantic segmentation is a computer vision task in which each pixel of an image is categorised into a class. It is used for a wide range of applications, including scene understanding, augmented reality, autonomous driving, video surveillance and anomaly detection. Semantic segmentation datasets, however, are difficult to create and maintain because they require meticulous labelling, which is time-consuming, expensive, and requires experts' assistance. Thus, datasets are often scarce, particularly in the medical and industrial fields, which is a problem since deep learning models for semantic segmentation tend to be data-

*. The authors would like to thank Cynthia I. Ugwu for her supervision and valuable input on this research

hungry as pointed out by Wu et al. (2023). Recent research has attempted to address this issue by replacing or extending datasets for semantic segmentation with synthetically generated datasets. This is achievable through a process known as semantic image synthesis. It aims at generating realistic images based on semantic masks provided as inputs like the work of Park et al. (2019); Sushko et al. (2022); Wang et al. (2022). Other works go even further and aim at generating both images and corresponding semantic masks like in Wu et al. (2023); Cechnicka et al. (2023); Shao et al. (2023); Macháček et al. (2023); Han et al. (2023); Fernandez et al. (2022); Zhang et al. (2021).

The approaches listed above represent a contrast to conventional computer vision approaches. Traditionally, the computer vision community has focused on engineering better network architectures to improve the performance on a fixed dataset. In recent years there has been a shift in research toward improving the data rather than designing complex network architectures. Therefore, this paper addresses semantic image synthesis via Diffusion Models (DMs). We improved the training and sampling strategies of the Semantic Diffusion Model (SDM) by Wang et al. (2022) and conducted extensive experiments using three famous semantic segmentation models. Our overall framework, dubbed Improved-SDM (ISDM), was applied to a small-scale industrial dataset and has proven the performance gain of introducing synthetic data during the training of the selected models. Our contributions can be summarized as follows:

- We generate synthetic datasets for semantic segmentation using Diffusion Models starting from a small-scale industrial dataset for quality control.
- We conduct extensive experiments to investigate the impact of introducing synthetic data during the training process of three famous state-of-the-art segmentation models analyzing the improvements and degradation of incrementally augmenting the percentage of synthetic data during training.
- We demonstrate that under certain dataset configurations, some models trained only on synthetic data can reach comparable results to cases trained with only real data. This result could be particularly valuable for the medical domain and other areas where strict privacy rules apply.

2 Related Works

Park et al. (2019) equipped GANs with their proposed conditional normalisation technique called SPADE. SPADE is an acronym for "spatially-adaptive (de)normalisation" layers that can better preserve the semantic information in the mask. Sushko et al. (2022) built their OASIS model on top of the implementation by Park et al. (2019) through redesigning the discriminator to obtain a U-Net-like semantic segmentation model. The feedback of this discriminator was no longer binary as for a conventional discriminator but on a pixel level. Wang et al. (2022) introduced Semantic Diffusion Models (SDMs), one of the first approaches using DMs for the task of semantic image synthesis. In SDM the noisy image serves as input for the encoder of the network, whereas the semantic mask, which is used to condition the diffusion process, is only fed into the decoder. Furthermore, the mask is forwarded through SPADE layers. The Latent Diffusion Model (LDM) by Rombach et al.

(2022) operates in the latent space reducing the computational costs. The integration of cross-attention layers makes LDMs highly flexible in conditioning on various inputs such as text or semantic masks.

3 Dataset

Some important datasets for semantic segmentation are Cityscapes (Cordts et al. 2016), ADE20K (Zhou et al. 2017), CelebAMask-HQ (Lee et al. 2020), COCO-Stuff (Caesar et al. 2018) and Pascal VOC 2012 (Everingham et al. 2012). However, most of these datasets consist of a large number of different real-world pictures and do not aim for anomaly detection. One of the best-known public datasets for industrial inspection is MVTec by Bergmann et al. (2021), which is however designed for unsupervised anomaly detection. Thus, it was necessary to use a less explored dataset. The Lemons Quality Control Dataset by Adamiak (2020) consists of 2690 images of lemons with defects labelled on a pixel level for fruit quality control. The dataset features nine different defect categories, seven of which concern actual product quality issues. Further details about the dataset are described in Appendix A. There is a considerable discrepancy between the frequency of the defect categories. We decided to focus on the most frequent classes, which contribute to more than 5 % of all defective pixels. This choice was also necessary due to time constraints as the thorough experimentation required the repeated training and evaluation of several deep learning models. Therefore, we also started from the simpler binary segmentation task before considering the multi-class scenario. We created three sub-datasets:

- Binary Blemish: "Blemish" is a collective class for several issues on lemons. In total, there are 200 images.
- Binary Blemish-Illness-Mould: to overcome the issue of limited availability of images we took the three most representative classes ("blemish", "illness" and "mould") and treated them as a single class. The resulting dataset consists of 790 images.
- "Multi-Class Blemish-Illness-Mould Dataset": it is composed of the three most frequent defect classes but considered separately with a total of 790 images.

4 ISDM: Improved Semantic Diffusion Model

We aim to explore the capability of DMs for semantic image synthesis on a small-scale industrial dataset. The main goal is not to generate images that look as realistic as possible but to find an approach to improve semantic segmentation models for vision-based anomaly detection. In particular, we make use of the SDM implementation by Wang et al. (2022) and adjust the training and sampling strategies which result in improvements on the tested datasets. Therefore, the proposed method is referred to as Improved-SDM (ISDM). The SDM implementation essentially represents a three-stage procedure. The training stage where a SDM is trained with typical configurations for DMs in general: $T = 1000$ diffusion steps, a linear variance schedule and the prediction of the mean as well as the standard deviation of the random noise, which is why L_{hybrid} is used as a loss function. The optional finetuning stage where the SDM is trained further with a nearly identical configuration except for an additional drop rate parameter that indicates the probability with which the

diffusion process is not conditioned on the semantic masks. Finally, the sampling stage where a synthetic dataset is generated from the trained SDM.

Our ISDM implementation stays with the original default parameters for training. However, we adapt the training algorithm to monitor the training epochs and their losses. The model checkpoint is only overwritten when the training loss improves. Although this cannot avoid possible overfitting, usable models can be trained consistently with this method. In addition, the time required to train usable models has been reduced by around half and the amount of disk space required has also been reduced. The second change to the original SDM implementation brings significant advantages in terms of runtime as well. One of the disadvantages of the SDM implementation is that for sampling, as for training, 1000 diffusion steps are used as default. This leads to long sampling times and high resource consumption. Since the implementation uses L_{hybrid} as a loss function and is based on Guided Diffusion by Nichol and Dhariwal (2021), it is possible to rescale timesteps during sampling. Therefore, we were able to reduce the diffusion steps to $T = 300$. These improvements are illustrated in more detail in Appendix B with visual examples.

We test our approach with three well-known semantic segmentation models to have a proper estimation of the improvements and/or deterioration of the introduced synthetic data. In particular, we follow the evaluation approach by Macháček et al. (2023) and use the U-Net++ by Zhou et al. (2018), the FPN by Lin et al. (2017) and the DeepLabv3+ by Chen et al. (2018). All models feature a ResNet34 encoder pre-trained on ImageNet. They are trained for 50 epochs, with an image resolution of 256×256. There is no data augmentation applied during training. The training process is monitored using a validation split of 20 % from the real training set. After training the models are evaluated on the real test set. For the evaluation metric, we use the Intersection over Union (IoU) widely used in semantic segmentation. It calculates the ratio between the intersection and the union of the ground truth and predicted semantic masks and is defined as:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{1}$$

where TP, FP, TN and FN are the numbers of true positive, false positive, true negative and false negative predicted pixels respectively. It should also be noted that the aggregations of the IoU score over images and classes are often simply referred to as mean IoU (mIoU).

For any given dataset and segmentation model, the proposed evaluation process first trains and evaluates the model on the real data only. This model serves as a baseline for comparison. Further models are trained using varying dataset configurations, including both real and synthetic data. The percentage of synthetic data is gradually increased to verify if adding too much synthetic data could lead to the deterioration of the model performance.

5 Results

The following sections report the results of the proposed method on the binary and multi-class sub-datasets. The different dataset configurations are expressed in terms of numbers on real #R and synthetic #S data used to train the downstream segmentation model. For the values of #S below 1000, a random subset of the 1000 generated image-mask pairs is chosen. The training and evaluation of each segmentation model on each dataset configuration was

repeated three times to exclude possible coincidences during training. We report the average mIoU from these three runs and compare it to the mIoU of the baseline, which is represented as black dashed line in the following figures.

5.1 Binary Blemish Dataset

The results of the experiments with the ISDM method on the Binary Blemish Dataset are reported in Figure 1. It emerges that nearly all of the tested configurations could outperform the baseline with a significant improvement when using 400 generated image-mask pairs or more. For the U-Net++ and the DeepLabv3+, the models improved even further as more synthetic data was added, although the differences between the configurations with #S=700 and #S=1000 are marginal. For all models, it was also possible to train them with 1000 synthetic image mask pairs alone and achieve a mIoU score comparable to the baseline.

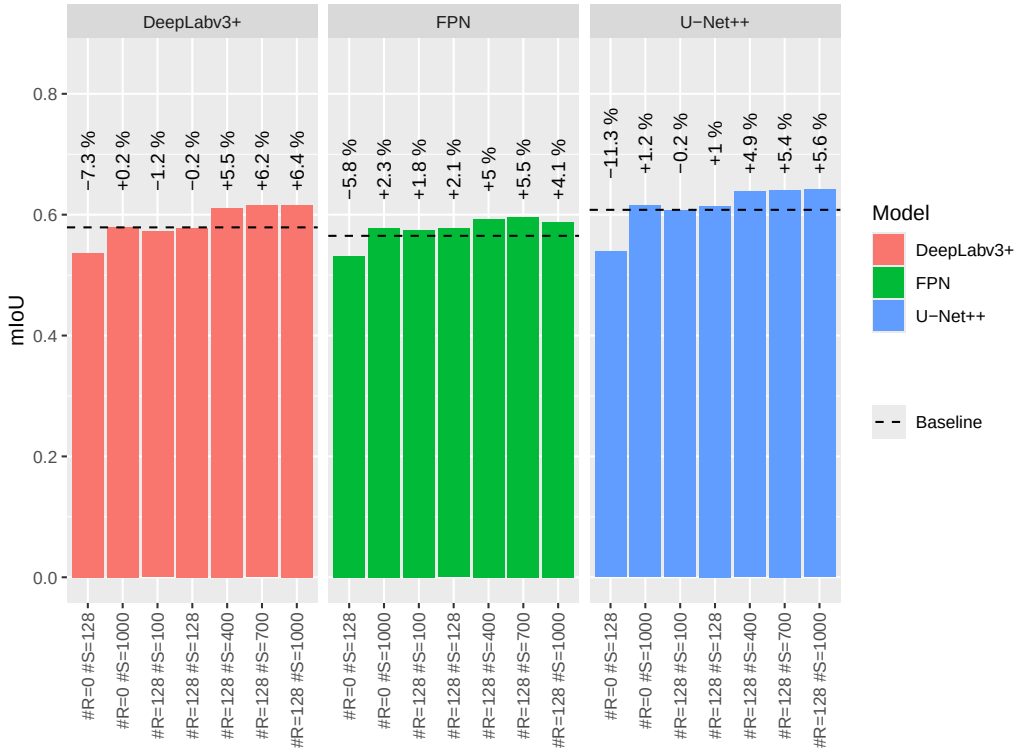


Figure 1: Results of the ISDM method on the Binary Blemish Dataset.

A shortcoming of the ISDM approach is that the models tend to overfit when trained on small datasets, however, the conditioning on the semantic masks worked so well that the defects are even more visible than in the original image (see Figure 9 Appendix C).

5.2 Binary Blemish Illness Mould dataset

The Binary Blemish Illness Mould dataset combines the three most common defect classes into a single one overcoming the problem of overfitting (see Figure 10 Appendix C). However, Figure 2 shows that the generated data could hardly achieve improvements in terms of mIoU

compared to the baseline. The best-performing configuration is #R=504 #S=1000 for all three models. Moreover, the data generated with the finetuned ISDM usually achieve higher improvements than the data generated without finetuning. The largest differences regarding finetuning can be observed when the segmentation models were trained on synthetic data only (#R=0). These differences amounted up to +6.4 %pt. for the U-Net++, +9.6 %pt. for the FPN and +7.7 %pt. for the DeepLabv3+ between the finetuned and non-finetuned models in comparison to the baseline.

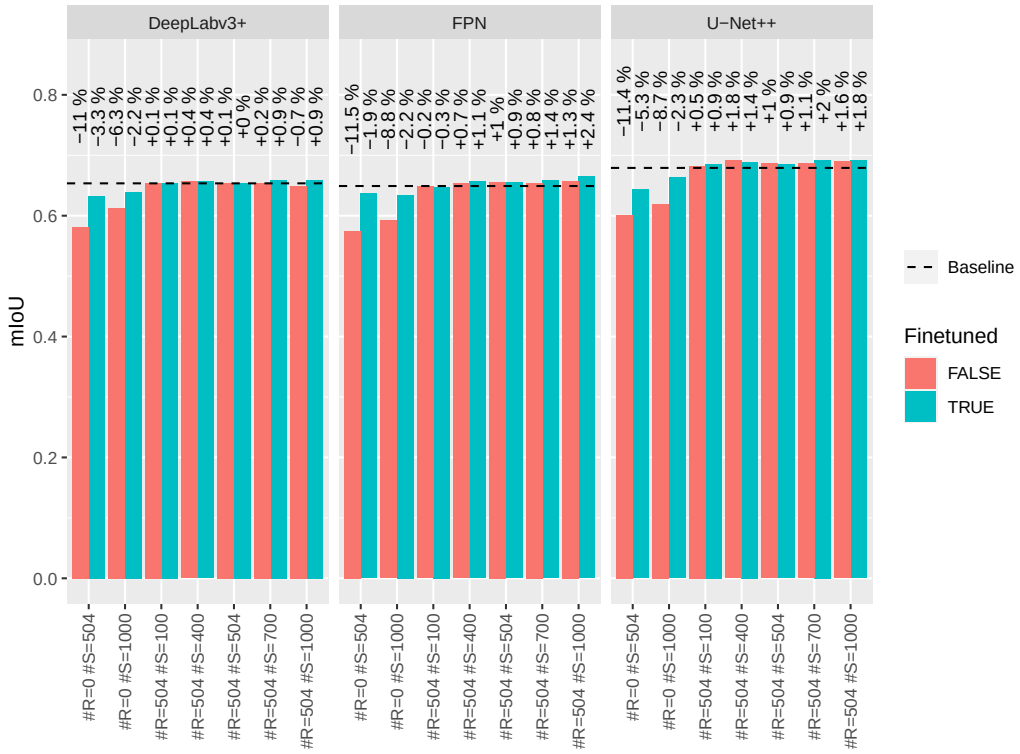


Figure 2: Results for the ISDM Method on the Binary Blemish-Illness-Mould Dataset for a Finetuned and a Non-Finetuned ISDM.

We investigated the possibility that better results could be achieved by using more synthetic data. Figure 3 shows that larger amounts of synthetic data help especially when the semantic segmentation models are trained only on synthetic data. The improvements for adding more synthetic data are smaller for the configurations with a mix of synthetic and real data. The configuration #R=0 #S=4000 could keep up with the baseline for all models. The configuration #R=504 #S=4000 proved to be the best for all models.

5.3 Multi-Class Blemish-Illness-Mould Dataset

The results for the Multi-Class Blemish-Illness-Mould Dataset are shown in Figure 4. There are improvements through the synthetic data for all three models and all three classes. However, the best dataset configuration for one class does not automatically prove to be the best for the other two classes. For instance, the DeepLabv3+ achieved the greatest im-

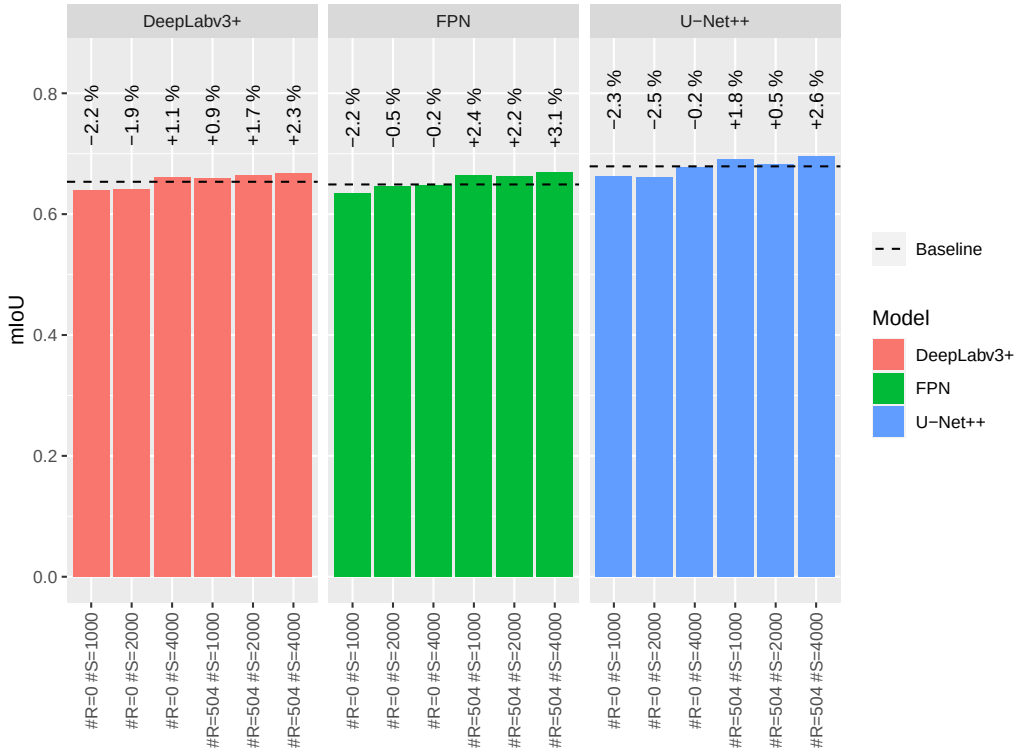


Figure 3: Comparison of Different Amounts of Synthetic Data on the Binary Blemish-Illness-Mould Dataset. The percentages refer to the improvements/deterioration compared to the baseline.

provement for "blemish" (+7.0 %) with the configuration #R=504 #S=700 which resulted in a decrease (-1.8 %) for the "illness" class. Likewise, the U-Net++ reached its best improvements with different configurations for each class. The FPN profited from the use of many synthetic image-mask pairs (700 or 1000) for "blemish" and "illness" but "mould" improved most when only 100 synthetic images were used. However, there were also configurations for each model with which all three classes improved, or at least did not deteriorate, compared to the baseline. Therefore, the configuration #R=504 #S=700 should be chosen for the FPN, as it gave the best average improvements across classes for this model. For the U-Net++ the configuration #R=504 #S=504 could offer the best compromise and for the DeepLabv3+ the configurations #R=504 #S=400 and #R=504 #S=700 were the only ones where adding the synthetic data did not degrade any of the classes.

6 Conclusion

Our work aligns with a recent paradigm shift in deep learning that was made possible through the advances of generative models: instead of fixing the dataset and attempting to improve the architecture of the neural network, the model is fixed and the dataset is enhanced through synthetic data. As semantic segmentation datasets are scarce, particularly in the medical and industrial fields, we used Diffusion Models to supplement existing

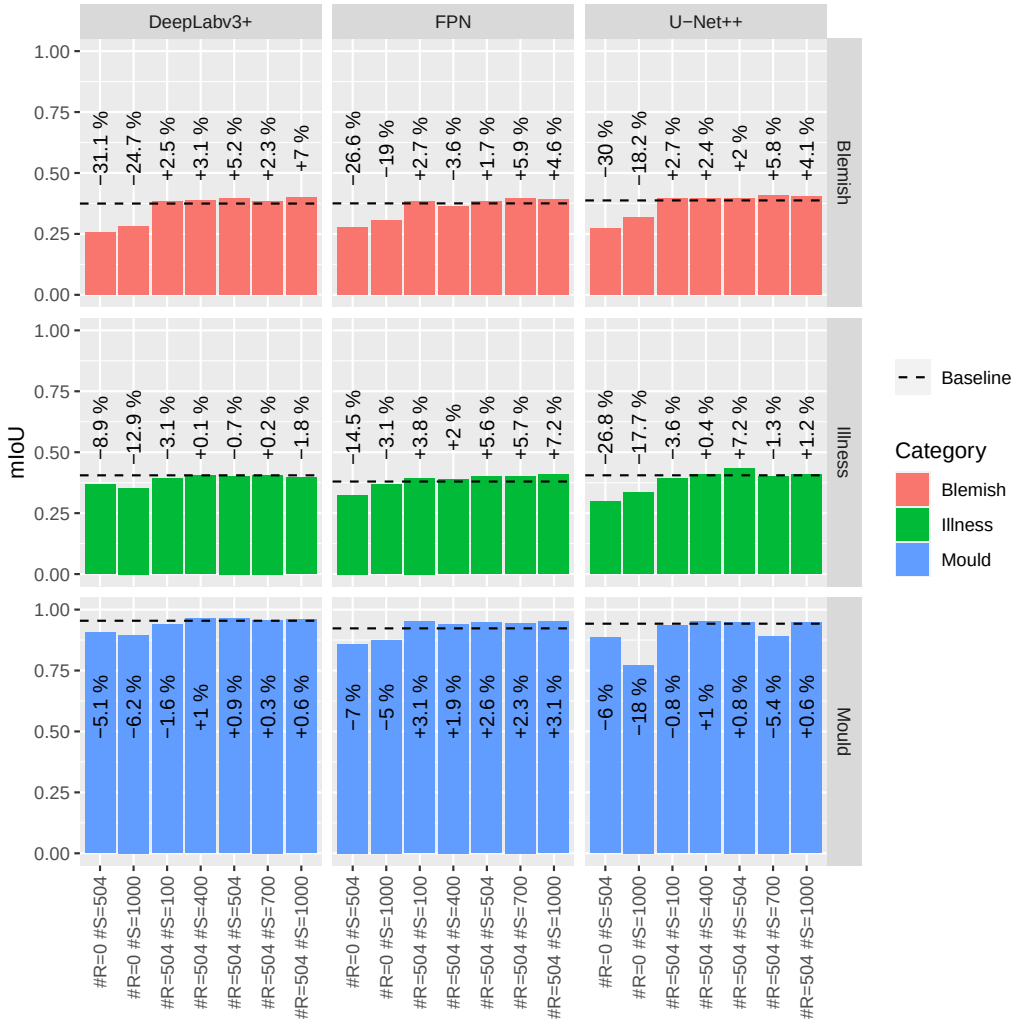


Figure 4: Results for the ISDM Method on the Multi-Class Blemish-Illness-Mould Dataset. The percentages refer to the improvements/deterioration compared to the baseline.

datasets with generated images. We developed ISDM to synthesize synthetic data for the industrial Lemons Quality Control Dataset. The impact of the generated data was tested on three well-known semantic segmentation architectures DeepLabv3+, FPN and UNet++. Extensive experiments have shown that ISDM was able to generate binary and multi-class segmentation datasets that improved all three models, although the amounts of the improvements differed between model architectures and dataset configurations. The most important insight is that for some models that were trained only on synthetic data, we could achieve scores comparable to the real baseline. This result could be particularly valuable for the medical domain and other areas where strict privacy rules apply.

References

Maciej Adamiak. Lemons quality control dataset, Jul 2020. URL `https://github.com/softwaremill/lemon-dataset`.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

Sarah Cechnicka, James Ball, Callum Arthurs, Candice Roufosse, and Bernhard Kainz. Realistic data enrichment for robust image segmentation in histopathology. *arXiv preprint arXiv:2304.09534*, 2023.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Marc Everingham, Luc Van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S Graham, Tom Vercauteren, and M Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 79–90. Springer, 2022.

Kun Han, Yifeng Xiong, Chenyu You, Pooya Khosravi, Shanlin Sun, Xiangyi Yan, James Duncan, and Xiaohui Xie. Medgen3d: A deep generative framework for paired 3d image and mask generation. *arXiv preprint arXiv:2304.04106*, 2023.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

Roman Macháček, Leila Mozaffari, Zahra Sepasdar, Sravanthi Parasa, Pål Halvorsen, Michael A Riegler, and Vajira Thambawita. Mask-conditioned latent diffusion for generating gastrointestinal polyp images. *arXiv preprint arXiv:2304.05233*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/nichol21a.html`.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Shitong Shao, Xiaohan Yuan, Zhen Huang, Ziming Qiu, Shuai Wang, and Kevin Zhou. Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. *arXiv preprint arXiv:2304.13416*, 2023.

Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. Oasis: only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision*, 130(12):2903–2923, 2022.

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023.

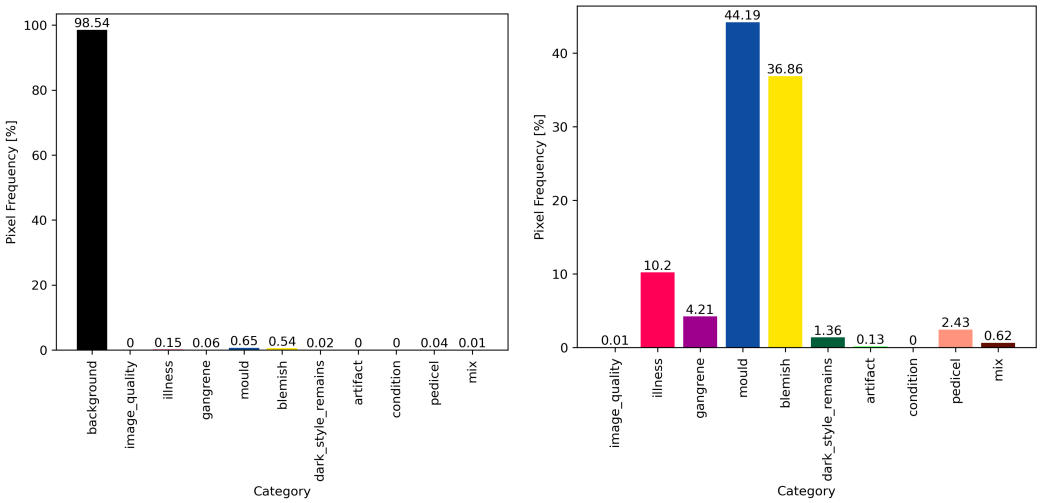
Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00889-5.

Appendix A. Lemons Quality Control Dataset

The Lemons Quality Control Dataset by Adamiak (2020) is one of the few publicly available datasets addressing semantic segmentation for anomaly detection in the industrial domain. As the name of the dataset, dubbed Lemons Dataset, suggests, it consists of 2690 images of lemons with defects labelled on a pixel level for fruit quality control. The dataset comprises nine different defect categories with a considerable discrepancy between their frequencies. The healthy tissue of the lemons is not labelled, therefore the background of the image and the healthy parts share a single label (with category ID 0) and are simply referred to as the "background" category. Figure 5 represents the distribution of the different defect categories with (5a) and without (5b) considering the "background" class. Figure 6 shows some example images and their corresponding ground-truth labels (semantic masks). These figures demonstrate that most of the pixels in the dataset belong to the "background" category, only approximately 1.5 % of the pixels are defective. Datasets with imbalanced and underrepresented categories generally decrease the performance of semantic segmentation models, which is why we decided to focus only on subsets of the Lemons Dataset.



(a) Distribution with "Background". (b) Distribution without "Background".

Figure 5: Histograms of the Pixel Frequency for the Categories of the Lemons Dataset. The "background" category comprises the real image background as well as the healthy tissue of the lemons and is therefore extremely predominant in terms of pixel frequency (left). To represent the distribution of the various defects on the lemons, the "background" pixels need to be ignored (right).

Appendix B. Improved Training and Sampling Procedures

We want to point out some differences introduced by our improved training and sampling procedures:

1. The introduction of a monitoring mechanism leads to the saving of model checkpoints that improve with further training steps.

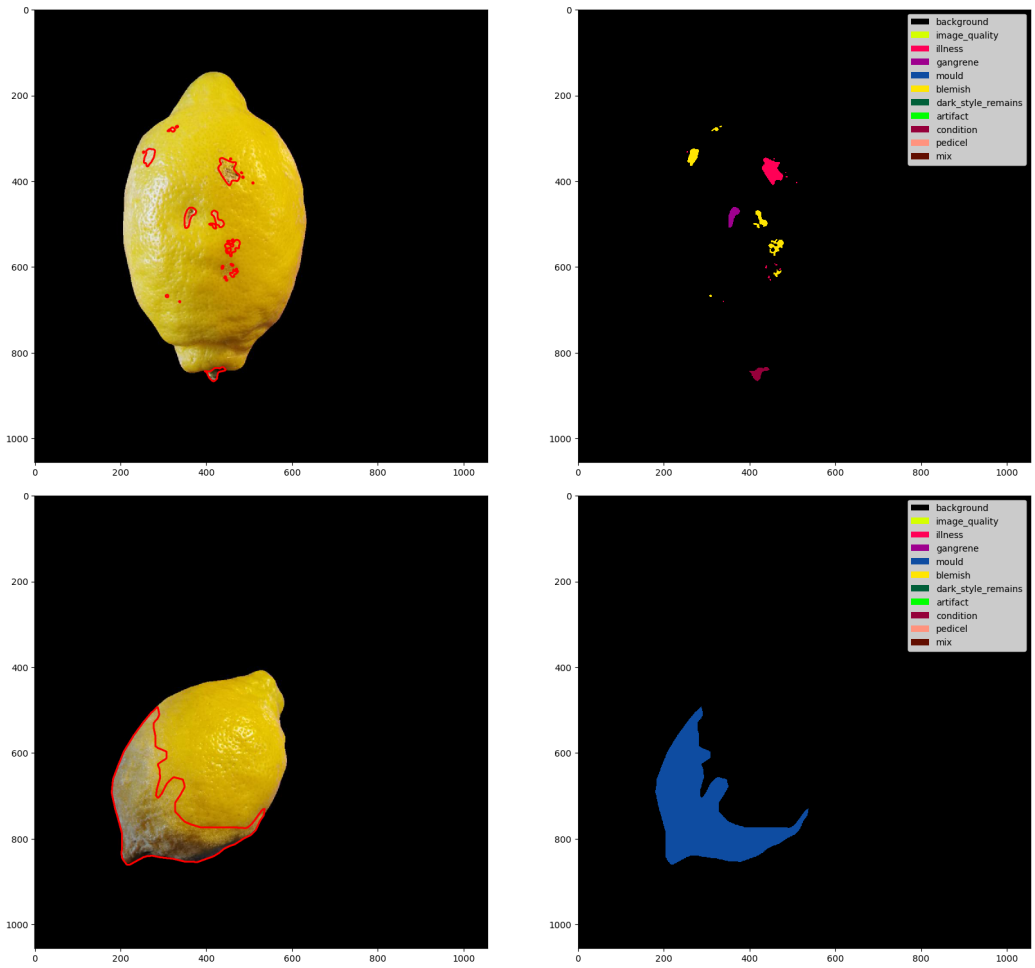


Figure 6: Examples from the Lemons Dataset. The images (left) are shown alongside the corresponding semantic masks (right). Each category of the dataset is mapped to a specific colour that can be looked up on the legends. The ticks on the axes show that the original image resolution is 1056×1056.

2. The use of timestep rescaling during sampling not only helps to reduce the necessary diffusion steps and thus runtime costs but also leads to better sampling results.

All three images in Figure 7 were sampled using the same semantic mask for conditioning. The first image, which appears the noisiest among all three, was sampled from an SDM trained for 29142 steps and using 300 diffusion steps and timestep rescaling for sampling. The second image was sampled with the same procedure from a later checkpoint (53664 optimisation steps) of the same model. It shows some noise as well but seems the best among all three. Note, that both of these checkpoints would never been saved without our modifications, previously, the training algorithm saved a checkpoint arbitrarily after each 10000 optimisation steps without considering any metric. The last image in the figure was sampled from the same model, but using 1000 diffusion steps for sampling without rescaling.

(a) Sample from ISDM Trained for 29142 Steps.

(b) Sample from ISDM Trained for 53664 Steps.

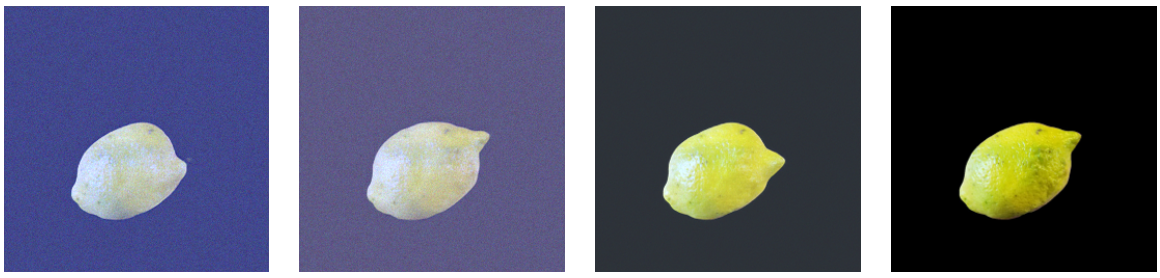
(c) Sample from ISDM (53664 Steps) without Rescaling.

Figure 7: Example for Three Differently Sampled Images. All three images were sampled from the same ISDM, but at different points in training, images (a) and (b), or with and without the rescaling of diffusion steps, images (b) and (c).

It seems better than the first, but worse than the second, although its sampling process lasts three times longer.

Appendix C. ISDM Shortcomings

We observed that the same ISDM, when conditioned with the same semantic mask, could produce samples of vastly different quality. This is illustrated in Figure 8 and is a reason why researchers, such as Shao et al. (2023), chose to design filtering mechanisms to exclude bad samples. So, it seems that problem is not entirely unknown but applies to other DMs as well. This problem could also be the reason, why ISDM could only achieve moderate and inconstant improvements.



(a) Sample 1.

(b) Sample 2.

(c) Sample 3.

(d) Real Image.

Figure 8: Example of Three Identically Sampled Images of Different Perceived Quality. The same ISDM and the same semantic mask were used for generating all three samples. The semantic mask belongs to the real image shown in (d).

Another observation is that using the small Binary Blemish Dataset resulted in overfitting, which is illustrated in Figure 9. Despite the clear overfitting, we validated the generated data and included the results in this work since the objective to improve the

baseline model was still being achieved. The problem of overfitting was overcome on the larger Binary Blemish-Illness-Mould dataset, Figure 10 shows that for the identical conditioning information, ISDM could sample two lemons that differ from each other in shape, colour and other features. They also differed significantly from their real image.

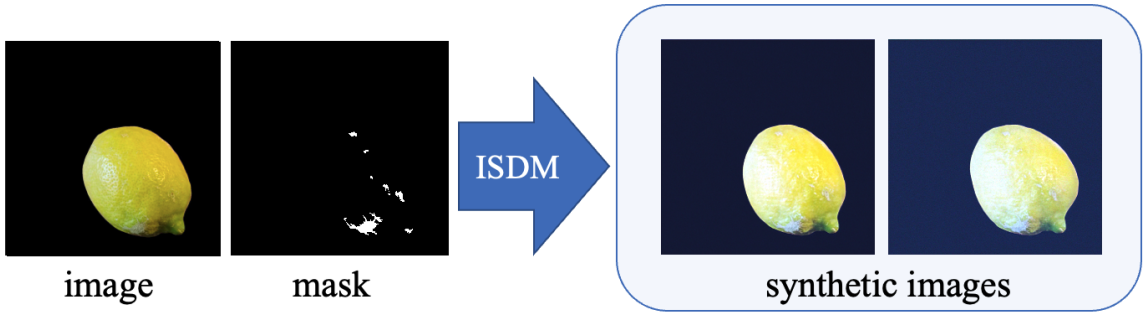


Figure 9: ISDM on the Binary Blemish Dataset. The model generates images that show hardly any changes in shape and size compared to the original.

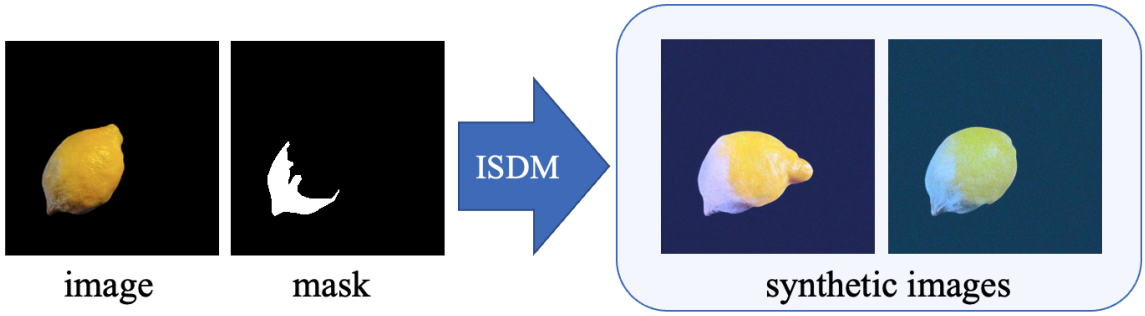


Figure 10: SDM on the Binary Blemish-Illness-Mould Dataset. The two synthetic images on the right show no signs of overfitting as the lemons differ in shape and colour.