# Enhancing Histology Datasets With Synthetic Data for Semantic Segmentation

Bc. Tomáš Tánczos*
*Supervised by: prof. Ing. Vanda Benešová, PhD.*†

Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava / Slovakia

## Abstract

Analyzing digital histopathological images is crucial in medical diagnostics; however, obtaining large, well-annotated datasets is challenging. This work focuses on augmenting histopathological datasets using generative neural networks and evaluating the new data's influence on the deep learning-based segmentation model. The analysis examines current methods for generating synthetic images and compares them to those that best meet our requirements. Based on this evaluation, we decided to prioritize denoising diffusion probabilistic models over generative adversarial networks due to their ability to perform image synthesis and inpainting. Because of the nature of image synthesis from noise and image inpainting processes, our solution combines these two and leverages their potential in dataset augmentation. The proposed solution experiments on in-house histopathological image Figure 3 datasets of heart tissue because, during the previous research, the blood vessel class showed a significant underrepresentation. We experimented with image synthesis in pixel and latent space with the same model architecture to test if we could capture better features with latent representation. The paper evaluates the quality of generated images using quantitative metrics and visual analysis. The high-level overview of our work is visible in Figure 1. Our synthetic dataset improved the DICE score of blood vessels in the segmentation model by 2 %.

**Keywords:** Dataset Augmentation, Generative Neural Networks, Semantic Image Synthesis, Digital Histopathology

## 1 Introduction

Analyzing digital histological images is crucial for patient treatment and diagnosis. Researching deep learning-based methods for analyzing digital histological images is a topical issue. However, collecting large numbers of well-annotated digitized histology data for the training of deep neural networks is difficult. Creating a histology dataset is more complex and time-consuming than labeling a general one because of the need for domain knowledge. Another aspect that makes creating a public medical dataset harder is the patients' privacy, which is classified as sensitive data. The lack of images is a limiting factor for research on new approaches using deep neural networks, as one possible solution is the creation of synthetical images to augment the dataset. The root cause of our research is the insufficient class representation in our in-house heart tissue dataset for semantic image segmentation, which causes insufficient segmentation performance, especially for blood vessels.

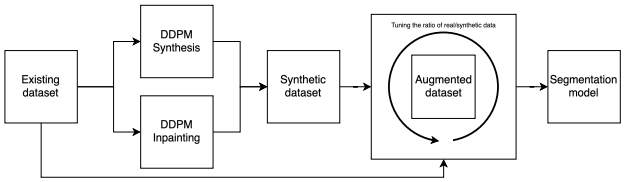


Figure 1: Overview of the dataset augmentation process using image synthesis and inpainting. Synthetic data is generated and combined with the original dataset to train a segmentation model.

**Our goal** is to enhance the existing segmentation model [4] for higher biological structures and examine the impact of dataset augmentation with fully or partially synthetic images. **Our solution** includes two crucial points. Firstly, we developed a deep learning-based method for semantic image synthesis. In this part of the work, we experimented with two approaches for creating fully synthetic images or partially editing them by inpainting. Both methods were performed with denoising diffusion probabilistic models (DDPM) in pixel and in latent space. The second step involves evaluating and comparing our augmented datasets with the baseline dataset. The synthetic dataset with the best results was used as an additional dataset for the image segmentation task to determine if the augmentation helped the model better learn underrepresented features. **Our contribution** is to create a histological dataset of heart tissue from synthetic data.

*tomas.tanczos@stuba.sk
†vanda_benesova@stuba.sk

## 2 Related works

During our research, we focused on two groups of papers that used histological data with DDPM and augmented the dataset with synthetic data.

### 2.1 Image inpainting utilization in digital histology

Various methods, such as folding or uneven illumination, can damage tissue slices during manipulation before they are digitized. These damages are referred to as artifacts, and they can make it challenging to analyze the images. Removing artifacts from WSI images is crucial in medical imaging but can be difficult. Several solutions exist based on GANs to remove artifacts from samples, but they risk the miss-transfer of stain style because they generate a whole image and not only the affected part. To address this problem, Zhenqi He et al. [6] proposed the ArtiFusion model utilizing DDPM Figure 2. The most crucial point of this paper is that they do not generate whole synthetic images as conventional DDPMs. However, in this case, their proposed model generates just part of the image to replace the artifact. They used an inpainting approach from Lugmayr, Andreas, et al. [8]. Contrary to DDPM, the second change is replacing the U-Net architecture with a novel Swin-Transformer-based network. Their network leverages attention mechanisms better to capture both local and global relationships in histology images, improving restoration quality.


Figure 2: Artifact removal comparison of CycleGAN approach against inpainting with DDPM (ArtiFusion). The GAN modifies the whole image, contrary to ArtiFusion, which focuses only on the artifact [6].

### 2.2 Datasets augmentations with synthetic images

In [10], the aim of Mathias Öttl et al. was to achieve a better performance of the U-Net segmentation model by augmenting the dataset with synthetic images. They work with breast cancer tumors and are focused on Human Epidermal growth factor Receptor 2 (HER2) and its subtypes. Usually, the HER2 comprises several subclasses, and the final treatment is based on this combination. However, these subtypes can be imbalanced, and some can be underrepresented, leading to a weaker segmentation performance. The usual augmentation methods, such as class oversampling, can quickly turn into model overfitting, so the paper aimed to extend the dataset with subtype-balanced synthetic images generated by generative neural networks. The work compared three approaches: image generation with GAN, DDPM, and image inpainting with DDPM. All three models generated a new image based on existing masks in their baseline dataset; however, the class types of masks were mixed. Their experiments were performed on 40 pieces of whole slide images, split into train-validation-test sets of 24-8-8 [10]. In the end, the results were evaluated quantitatively and qualitatively. The experiment in which the original dataset was extended by 100 % of synthetic images generated by DDPM was marked as the best. There, the Dice score of the U-Net segmentation model increased by 2.43 % to 0.854. From a qualitative perspective, they reported that images from all three models are visually very close to the real samples. The GAN generated repeating samples, a common problem with these networks. Images from DDPM had a higher variance and small artifacts in the background. The inpainted images also showed a high variance in subtypes, but the staining on the new image was more similar to the original image [10].

As mentioned earlier, we need a large amount of labeled data to successfully apply deep learning methods to solve these tasks, which is time-consuming and requires an expert pathologist. To overcome this challenge, Xinyi Yu et al. [16] proposed a two-stage synthetic image generator based on DDPM [7], with the goal of dataset augmentation. The reason for not choosing GAN architecture was their unstable training and low variance in generated images [16]. The output of the first step of their proposal is a nuclein instance map generated by unconditional DDPM. The work did not discuss the exact architecture of the U-Net used in this step. The resulting instance map is used as additional information and inspired by Wang et al. [15], they embed it into the network through SPADE [11]. The network used to generate the final synthetic nuclei image in the second step is fully adapted from [15]. Two distinct nuclein segmentation datasets from multiple organs were used for the experiment, with 44 and 30 whole-slice images. The authors stated that the final images look realistic and diverse. They also underlined that the generated images are well-aligned with a synthetic nuclein structure, a crucial requirement in segmentation training. In experiments, two models (Hover-Net and PFF-Net) were trained to segment nucleins, and the Dice coefficient and Aggregated Jaccard Index were used as evaluation metrics to quantify their results. The experiments with synthetically extended datasets were evaluated on both mentioned datasets. The work highlighted that even adding 10 % of augmented samples can perform better than without synthetic data [16].

## 3 Proposed solution

We will describe the dataset used for experiments and our overall inference pipelines for the solution, covering image synthesis and inpainting.

### 3.1 Our dataset

This work focuses on whole slide image (WSI) data acquired from the Institute for Clinical and Experimental Medicine (IKEM) in Prague[1], an in-house dataset. This dataset was previously utilized for segmentation in [4], which we aim to support through our research. The images from IKEM show heart tissue biopsies taken after heart transplantation (Figure 3.) and include various higher-level biological structures such as blood vessels, inflammations, and endocardium. At the start of the work, we had access to 51 fully or partially annotated images; however, later, we were provided with other samples containing only blood vessels, which we also incorporated. Each image has a resolution of approximately $10,000 \times 10,000$ pixels. In addition to these images, annotations of higher-level biological structures are provided in GeoJSON format. We have annotated the following classes: endocardium (31.22 %), inflammation (38.40 %), blood vessels (8.14 %), and fatty tissues (22.03 %). The dataset's preprocessing process was performed as follows. The primary goal was to extract image patches of a satisfactory size of 256 pixels to serve as input for our models. We went through the images systematically with window size of 128 pixels horizontally and vertically. The patch was kept if a given criteria was fulfilled — it contained at least 50 % tissue and 10 % of some label.

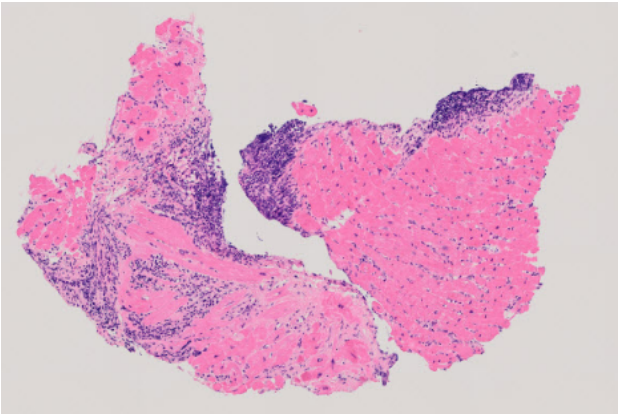

Figure 3: A real-world image sample of a whole slide image of heart tissue biopsy provided by IKEM. The image is captured with a medical scanner and colored with H&E staining. The figure size is more than 10,000 x 10,000 pixels, and a special application is needed for its detailed examination.

### 3.2 Image synthesis

Figure 4 illustrates the inference process of our proposed solution. The right side (red) of the figure highlights the creation of fully synthetic images, where the synthesis model has a Gaussian noise and corresponding semantic masks on the input, based on which the synthetic image will be created. On the left side (blue), the inpainting process is illustrated, where the model gets on the input semantic mask again; however, instead of noise, we have a partially covered image where we want to inpaint based on the semantic mask. In the middle of the image is the synthesis model, which has an optional Vector Quantized Variational Autoencoder [14] part used for synthesis in latent space. We trained two models, one for pixel space generation and one for latent space. The synthesis models are trained as a traditional DDPM [7], but thanks to the nature of the inpainting process [8], we can use this model during the inference in both ways without any other modification.

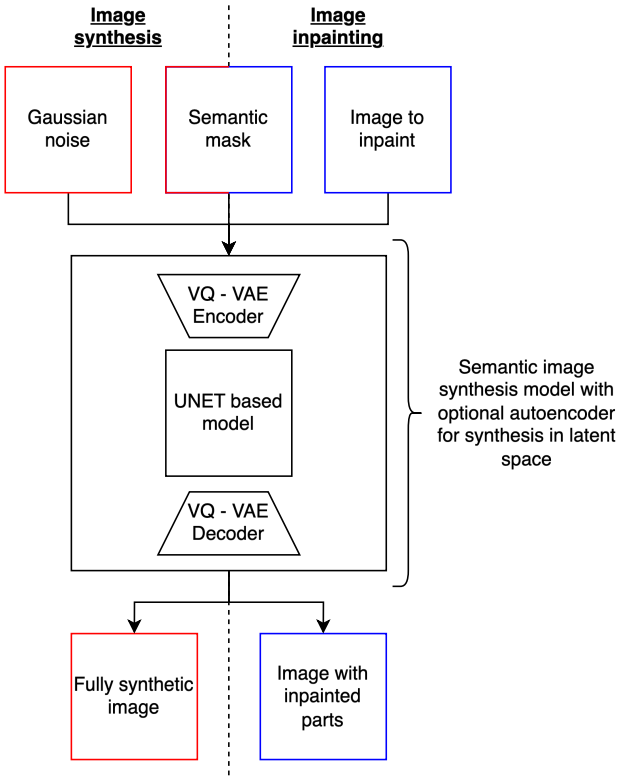


Figure 4: Inference process of our semantic synthesis model, the right side (blue) presents the fully synthetic image synthesis and the left side (red) stand for image inpainting.

Our approach uses an U-Net-based model for noise estimation Figure 5 to follow a traditional encoder-decoder architecture tailored for image processing. The encoder is implemented as a ResNet-based [5] encoder, which extracts hierarchical features using residual connections [5]. For the decoder Figure 6, we adapt the approach from [15]; the decoder is developed to reconstruct the original image resolution while embedding additional context from the encoder and the segmentation map. A key enhancement in the decoder is the integration of spatially adaptive denormalization (SPADE) layers [11]. These layers spatially modulate the feature maps based on the segmen-

[1] https://www.ikem.cz/cs/

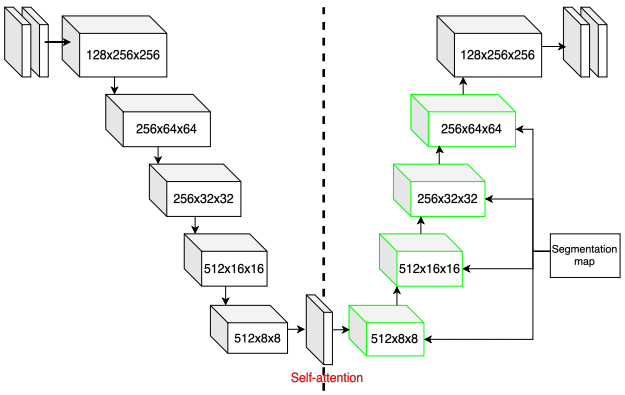


Figure 5: U-Net based noise estimator for our image synthesis process, with self-attention in botlleneck.

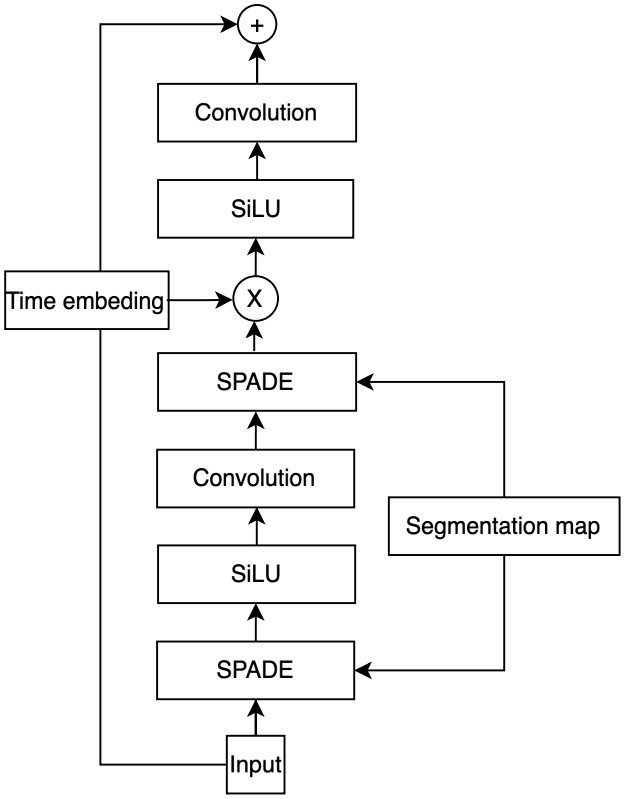


Figure 6: Detailed view of decoder block for noise estimator.

tation map, allowing the model to adaptively align reconstructed features with the semantic information. SiLU [12] activation functions in the network blocks are used to introduce smooth, non-linear transformations that improve gradient flow in deeper models.

The Autoencoder is a vector quantized variational model [14] designed to encode input images into a discrete latent space and reconstruct them, enabling the use of the Latent Diffusion Models (LDMs) approach [13]. The encoder blocks employ residual connections and attention mechanisms for feature extraction. It concludes with a vector quantization step, replacing continuous latent vectors with discrete representations from a code book. The decoder mirrors the encoder to up-sample the latent representation and reconstruct the original resolution progressively. It incorporates attention mechanisms to guarantee semantic coherence.

### 3.3 Semantic image segmentation

We used a ResNet-based [5] U-Net architecture for semantic image segmentation to segment the biological structures in our heart tissue images. The model follows an encoder-decoder structure, with the ResNet blocks extracting hierarchical features from the input images.

To improve the feature selection in the decoder, we integrated an Attention Gate (AG) mechanism introduced in [9] into our U-Net-based segmentation model. The attention mechanism is developed to refine the spatial feature maps by selectively suppressing irrelevant activations while highlighting the most informative regions. Given a gating signal $g \in \mathbb{R}^{C\times H\times W}$ from the decoder and skip connection features $x \in \mathbb{R}^{C\times H\times W}$ from the encoder, the AG computes an attention map $\alpha \in [0,1]^{1\times H\times W}$ as follows:

$$g' = W_g g + b_g \tag{1}$$
$$x' = W_x x + b_x \tag{2}$$
$$\psi = \sigma\left(W_\psi \cdot \mathrm{ReLU}(g' + x') + b_\psi\right) \tag{3}$$
$$\alpha = \mathrm{Upsample}(\psi) \tag{4}$$

Here, $W_g, W_x, W_\psi$ are learnable weights, $\sigma$ denotes the sigmoid activation, and 'Upsample' is used to match the spatial resolution. The final output is the element-wise multiplication of the attention map with the original skip connection:

$$\tilde{x} = \alpha \odot x \tag{5}$$

The model is trained using a combination of Binary Cross-Entropy and Dice loss. During training, it learns to predict pixel-wise class labels for each image, and its performance is evaluated using the Dice score and F1 score.

## 4 Experiments and results

Our experiments and results can be divided into two leading groups. The first group consists of sampling various synthetic datasets for augmentation. In the first round, we trained two synthesis models, one for sampling in pixel space and one for sampling in latent space. With both models, we created fully synthetic and inpainted datasets consisting of 10,000 samples and only with the class of blood vessels, which are the main targets of our interest. As mentioned in Section 3.1, we collected a new dataset after beginning our experiments. We retrained both models with this additional data to evaluate whether it improved their image synthesis performance. In the end, we

had six datasets (the inpainting samplings in pixel space in the time of paper writing were not done) that were evaluated quantitatively.

The second group of experiments contains a multiple training process for the segmentation model, where we gradually augment the baseline dataset with the best synthetic dataset from our first round of experiments. Table 1 describes the percentual class representation over the augmentation.

### 4.1 Experimental setup

This section presents our experimental setup, including the hardware and model hyperparameters. The synthesis model was trained on an NVIDIA RTX 6000 Ada Generation graphical card with 48 GB memory. We experimented with a multiple-model configuration, changing the levels in the U-Net architecture and testing self-attention applications at different levels. Finally, the best result was achieved with a model architecture presented in Figure 5. We down-sample the spatial dimension four times, and a self-attention layer is used only in the bottleneck of the model. The detailed description of the model blocks was provided in Section 3.2. With this configuration, the model has around 90 million learnable parameters. In our experiments, the diffusion process was configured with the following parameters: the number of steps in the denoising process was set to 1000, with a linear beta schedule running from 0.0001 to 0.02. We set the number of epochs to 500 with a batch size 16 and a learning rate 0.0002 for training. Input image values were scaled to the range [-1, 1] before input into the model. The optimization was performed using the Adam optimizer, with the learning rate specified above. The loss function used to train the diffusion model is the Mean Squared Error (MSE), as defined in eq. 6. The MSE loss is then computed between the ground truth noise and the predicted noise:

$$\mathscr{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_i - \hat{\varepsilon}_i)^2, \tag{6}$$

where $\varepsilon_i$ is the true noise added to the $i^{\text{th}}$ sample, and $\hat{\varepsilon}_i$ is the predicted noise.

The segmentation model was trained on an NVIDIA GeForce GTX 4090 graphical card with 24GB memory. The used model and loss function were discussed in Section 3.3. Every training session was initialized with a batch size of 16 and for 100 epochs. The learning rate was set to 0.0002 with the exponential scheduler to gradually reduce the learning rate during training and improve convergence stability [3]. Overall, we trained a five-segmentation model with dataset setup described in Table 1.

### 4.2 Results of semantic image synthesis

In this section, we evaluate the generated synthetic datasets against each other quantitatively. The results of our experiments are summarized in Table 2, which compares performance metrics for image generation in pixel and latent spaces. For evaluation, we used three metrics: KID, FID, and LPIPS [1, 2]. Both types of synthetic datasets were compared with our baseline dataset. The table shows that the datasets sampled in latent space performed better overall than images from pixel space. Interestingly, a retrained model with additional data in pixel space provided worse results than its base variation. The values within the group of latent sampling are very close to each other, especially for LPIPS; this could be a sign that our starting dataset was enough for our model, and the retraining does not provide much more new information. Finally, based on values from KID and FID, we selected the inpainted dataset sampled in latent space with the retrained model as best and will continue with it further in our experiments.

In Figure 7 we are providing a sample for inpainting blood vessels into a clear tissue with no other biological structure. In the third column, it is visible that the highlighted areas for the painting are properly modified, and the other parts of the image are untouched. No sharp transitions are visible on the inpainted areas, and the modification has a gentle border. However, determining if the inpainting is properly done and corresponds to a real blood vessel could not be told without an expert review, which will be covered in future work.

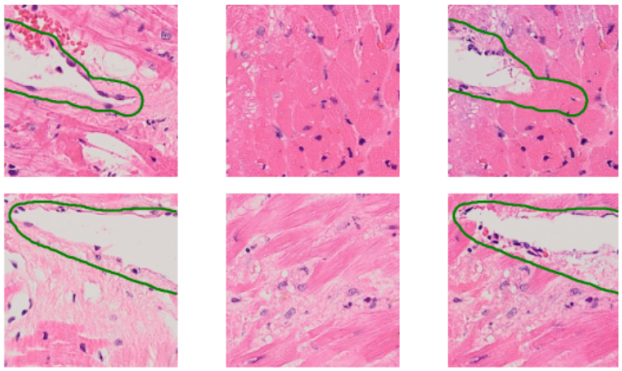

Figure 7: Sample inpainted images. The first column shows the original image with the related segmentation mask, the second column displays the tissue without other structures, and the third column presents the inpainted blood vessels.

### 4.3 Results of image segmentation

We separated three whole slide images for testing purposes, which were not included in the training dataset for the segmentation model. The summarized results are visible in Figure 8, which describes how the dice score for classes and F1 Score evolve with dataset augmentation. The dice score calculated here is the average dice for all three images. However, we augmented only the blood vessels class. It also affected the other classes. The best run

Table 1: Class representations percentage over the augmentation

| Dataset Variation | Endocard (%) | Blood Vessel (%) | Inflammation (%) | Fatty Tissue (%) |
|---|---|---|---|---|
| Baseline Dataset | 31.22 | 8.14 | 38.62 | 22.03 |
| Base + 2.5K Synthetic pcs. | 30.17 | 11.22 | 37.32 | 21.29 |
| Base + 5K Synthetic pcs. | 29.20 | 14.06 | 36.13 | 20.61 |
| Base + 7.5K Synthetic pcs. | 27.54 | 18.97 | 34.07 | 19.43 |
| Base + 10K Synthetic pcs. | 27.47 | 19.16 | 33.98 | 19.38 |

Table 2: Evaluation metrics in pixel and latent space for inpainted and synthetic data. ↓ indicates that lower values are better, and ↑ indicates that higher values are better.

| Synthetic dataset name | KID (↓) | FID (↓) | LPIPS (↑) |
|---|---|---|---|
| **Pixel Space** | | | |
| Fully Synthetic (Base Dataset) | 0.06 | 68.94 | 0.59 |
| Fully Synthetic (Extended Dataset) | 0.08 | 85.85 | 0.58 |
| **Latent Space** | | | |
| Fully Synthetic (Base Dataset) | 0.051 | 58.32 | 0.59 |
| Fully Synthetic (Extended Dataset) | 0.058 | 61.49 | 0.59 |
| Inpainted (Base Dataset) | 0.035 | 47.91 | 0.58 |
| Inpainted (Extended Dataset) | **0.03** | **47.61** | 0.58 |

appeared to be the one with the 7.5K pieces of data expansion, in which all metrics improved except for the endocardium.

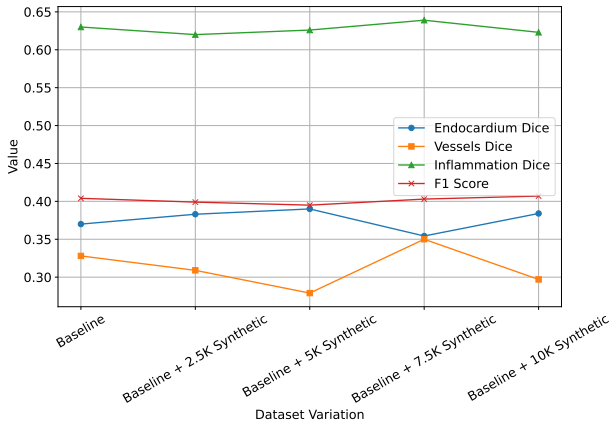


Figure 8: Evaluation metrics comparison across trained segmentation models. The chart displays Dice scores for different biological structures and F1 score across dataset variations.

We calculated the dice, precision, and recall metrics per image sample to understand the models' behavior better; this is summarised in Table 3. The results show that dice increased by around 4% for every sample. The data also shows that in the case of WSI with ID 7026, the augmented model was more conservative with higher precision and lower recall; it suggested fewer false positive values, but probably at the cost of missing true positives. This hypothesis is confirmed by Figure 9, where the red regions are for false positive predictions. However, blue regions stand for false negatives, and it is visible that in four cases, the model could not detect any part of the vessels. Green regions describe the true positives, and it is observable that the majority of the vessels were detected at least partially.

For the WSI with ID 3002, precision and recall increased as well. However, precision is still low. The low precision indicates false positive detections, also visible in Figure 9; here, the tissue has a leaky structure, and these holes probably confuse the model and are marked as blood vessels.

Table 3: Performance Comparison of segmentation model trained on baseline dataset and on additional 7.5K pieces Synthetic Data

| Test figure # | Dice | Precision | Recall |
|---|---|---|---|
| **Baseline dataset** | | | |
| #1 | 0.560 | 0.663 | 0.483 |
| #2 | 0.145 | 0.230 | 0.105 |
| #3 | 0.300 | 0.243 | 0.394 |
| **Base + 7.5K Synthetic pcs.** | | | |
| #1 | 0.529 | 0.810 | 0.393 |
| #2 | 0.194 | 0.214 | 0.177 |
| #3 | 0.343 | 0.254 | 0.528 |

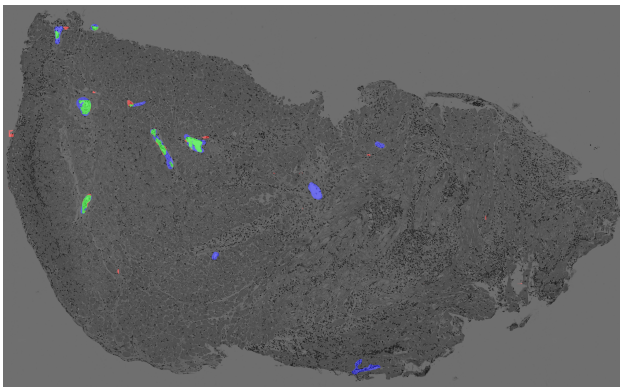

Figure 9: Comparision of True Positive (green), False Positive (red) and False Negative (blue) detections from the segmentation model trained with 7.5K pieces of synthetic data augmentation.

## 5 Future work

Our future work will consist of two parts. We have confirmed our hypothesis that augmenting the dataset with a

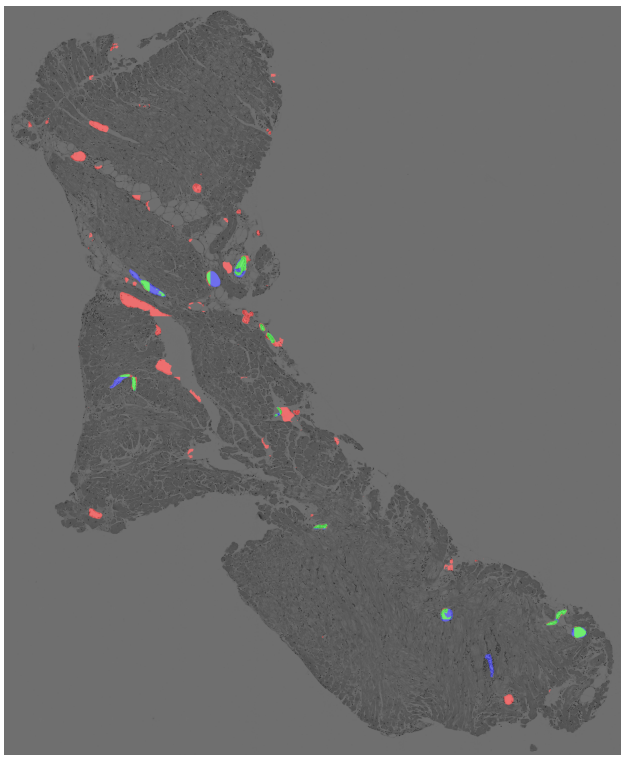

Figure 10: Detailed view of false positive (red) detections in the leaky heart tissue structure.

given amount of synthetic data can improve the segmentation model's performance; however, balancing precision and recall may require further experiments or a postprocessing step. The next step is applying this synthetic dataset to the model used in [4], which is specially developed for this dataset and more sophisticated than the one we used in segmentation experiments, and verifying our hypothesis again. We also want to verify our solution on other histological in-house datasets and support the development of deep learning solutions on them. Additionally, we plan to execute expert validation with pathologists to assess the clinical quality of our synthetic data.

## 6 Conclusion

In this work, we analyzed the influence of augmenting histopathology datasets with synthetic images on the performance of a deep learning-based segmentation model. Our approach leveraged denoising diffusion probabilistic models (DDPM) to generate fully synthetic and inpainted images, focusing on improving the segmentation of underrepresented classes, such as blood vessels. We evaluated our synthetic dataset using image quality metrics (KID, FID, LPIPS) and selected the best-performing dataset for segmentation experiments. Our findings demonstrate that augmenting the dataset with synthetic data can enhance segmentation performance, mainly for underrepresented classes. The best results were achieved with a 7.5K pieces of synthetic data augmentation, leading to improvements in Dice and F1 scores across multiple test samples. However, the results also underlined a trade-off between precision and recall, where the model trained with synthetic data showed a more conservative prediction behavior with higher precision at the cost of recall. The results suggest that while synthetic data improves class representation, further tuning or post-processing may be necessary to balance these performance metrics. Our results demonstrate the potential of synthetic images for enhancing histopathology datasets, supporting the development of more robust deep learning-based segmentation systems.

## References


[1] Samah Saeed Baraheem, Trung-Nghia Le, and Tam V Nguyen. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865, 2023.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[3] Florinel-Alin Croitoru, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, and Nicu Sebe. Learning rate curriculum. *International Journal of Computer Vision*, 133(1):291–314, 2025.

[4] Matej Halinkovic, Ondrej Fabian, Andrea Felsoova, Martin Kveton, and Wanda Benesova. Intrinsically explainable deep learning architecture for semantic segmentation of histological structures in heart tissue. *Computers in Biology and Medicine*, 177:108624, 2024.

[5] 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.

[6] Zhenqi He, Junjun He, Jin Ye, and Yiqing Shen. Artifact restoration in histology images with diffusion probabilistic models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 518–527. Springer, 2023.

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

[8] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[9] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[10] Mathias Öttl, Jana Steenpass, Matthias Rübner, Carol I Geppert, Jingna Qiu, Frauke Wilm, Arndt Hartmann, Matthias W Beckmann, Peter A Fasching, Andreas Maier, et al. Improved her2 tumor segmentation with subtype balancing using deep generative networks. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[11] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[12] Prajit Ramachandran, Barret Zoph, and Quoc Le. Swish: a self-gated activation function. 10 2017.

[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.

[14] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[15] 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.

[16] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer, 2023.