

Synthesis for Dataset Augmentation of H&E Stained Images with Semantic Segmentation Masks

Peter Sakalik¹, Lukas Hudec¹^a, Marek Jakab¹^b, Vanda Benešová¹^c and Ondrej Fabian^{2,3}^d

¹*Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovicova 2, Bratislava, Slovakia*

²*Clinical and Transplant Pathology Centre, Institute for Clinical and Experimental Medicine, Videnska 9, Prague 4, Czechia*

³*Department of Pathology and Molecular Medicine, 3rd Faculty of Medicine, Charles University and Thomayer Hospital, Videnska 800, Prague 4, Czechia*

Keywords: Medical data, Annotated Data Synthesis, Generative Adversarial Networks.

Abstract: The automatic analysis of medical images with the application of deep learning methods relies highly on the amount and quality of annotated data. Most of the diagnostic processes start with the segmentation and classification of cells. The manual annotation of a sufficient amount of high-variability data is extremely time-consuming, and the semi-automatic methods may introduce an error bias. Another research option is to use deep learning generative models to synthesize medical data with annotations as an extension to real datasets. Enhancing the training with synthetic data proved that it can improve the robustness and generalization of the models used in industrial problems. This paper presents a deep learning-based approach to generate synthetic histological stained images with corresponding multi-class annotated masks evaluated on cell semantic segmentation. We train conditional generative adversarial networks to synthesize a 6-channeled image. The six channels consist of the histological image and the annotations concerning the cell and organ type specified in the input. We evaluated the impact of the synthetic data on the training with the standard network UNet. We observe quantitative and qualitative changes in segmentation results from models trained on different distributions of real and synthetic data in the training batch.


1 INTRODUCTION


A histology tissue sample can be very easily abstracted as a non-stationary texture. The unique structures of a sample can be categorized as different texture classes. Therefore it is possible to apply similar methods to texture synthesis in the research on the synthesis of histology scans. Texture synthesis is the process of algorithmically creating a texture according to a small sample or a set of predefined characteristics. Depending on the applications, the algorithms must usually be both efficient and capable of generating high-quality outputs with high variability. The synthetic textures must be indistinguishable from the original sample or at least deceive the human observer. Over the years, plenty of texture synthesis methods have been introduced.


One of the simplest methods is random sampling, where the texture is sampled into tiles. They are then pseudo-randomly selected and joined together. However, the result is rarely sufficient as the seams between the tiles may remain visible.


Traditional approaches can be classified as pixel-based or patch-based depending on how large a sample is inserted into a new synthesized image. The most successful approaches use Markov arrays, non-parametric sampling, and tree-structured vector quantization. A new texture is formed by finding and copying pixels with the most similar neighboring pixels to the original texture. This technique limits the visible seams on the borders. However, the usual drawback is the repeatability of selected sampled patches.

Deep learning generative models have recently overcome traditional methods' generative quality. Extensive research has been done on generative adversarial networks (GANs) applications, which achieved significant results in various industrial and research fields. Their power breaks into more quality-requiring areas, such as the gaming industry and medicine.

^a <https://orcid.org/0000-0002-1659-0362>

^b <https://orcid.org/0000-0002-4329-6417>

^c <https://orcid.org/0000-0001-6929-9694>

^d <https://orcid.org/0000-0002-0393-2415>

One possible application is in pathology, which is the main focus of this paper. Structures of pathological findings are small and in large amounts, which makes it time-consuming for medical practitioners to annotate them perfectly. Automatic or semi-automatic annotation methods would be beneficial but introduce errors that produce noisy labels. Also, the semi-automatic methods usually require the initialization of parameters that may also take a significant amount of time.

The current cell segmentation datasets usually suffer from insufficient data quantity and variability because of the difficult annotation process. These datasets play a significant role in researching deep-learning models for cancer analysis, diagnosis, and staging.

This paper presents an automatic approach using GANs to generate many annotated synthetic data with quality similar to real samples. The generated annotated data can then be used for dataset augmentation necessary for training deep learning models, which makes models more robust, and better generalized. GANs reduce manual preparation time.

The contribution of this paper is the following:

- The presented method is specified for synthesizing hematoxylin and eosin (H&E)-stained histological images.
- The generated images are accompanied by annotation masks of cells of 4 classes.
- The model can generate visually organ-specific tissue and cells.
- Evaluation of the influence of synthetic data augmentations for semantic segmentation.

2 RELATED WORK

The quality of datasets depends on the accuracy of the segmentation masks. In most domains, they can be acquired by manual annotation or with semi-supervised segmentation methods. For this reason, different approaches have been developed to obtain them. Some diagnostic tools often provide a semi-supervised method for cell segmentation that can help the user/annotator with guidance or a set of parameters for automatic segmentation that can be later used for diagnostic analysis and support. One such software tool is QuPath (Bankhead et al., 2017), which uses simple thresholding combined with normalization and cell nucleus emphasizing. The user sets a set of threshold values that define the color interval of the hematoxylin purplish blue nucleus compared to

the pinkish eosin-stained extracellular matrix and cytoplasm. However, these simple segmentation methods may be insufficient for more complex images or cell classes. More advanced approaches are based on graph theory clustering methods or deep learning. Current deep-learning segmentation approaches achieve state-of-the-art performance. Specifically, these are the U-Net and U-Net++ architectures, respectively. Unfortunately, both of them require annotated images for training.

An alternative approach to acquiring annotated data is synthesizing tissue images where the research is open. As mentioned before, the histology tissue can be abstracted as texture, we present several state-of-the-art approaches for generating high-quality textures. The histology data are stored as the large resolution scans of the whole tissue, the Whole Slide Images (WSI), and a smaller scale specific selection, usually annotated in better detail, the Stain Tissue Microarray (TMA). There are known approaches to generate large-scale images, e.g., progressively growing GANs (Beers et al., 2018; Štepec and Skočaj, 2020) or specialized architectures for high-resolution image generation like StyleGAN (Karras et al., 2019), and BigGAN (Brock et al., 2018). The BigGAN is a network containing 355.7 million parameters with a generator output of 256×256 pixels. The StyleGAN, on the other hand, has only 26.2 million parameters and the generator output resolution is 1024×1024 pixels. Both architectures generate high-quality images.

Non-stationary texture synthesis with adversarial expansion (Zhou et al., 2018) presents a generative model that synthesizes texture by expanding the input sample from $k \times k$ to $2k \times 2k$ resolution. The architecture consists of a generator and two discriminators. The first one takes care of discriminating between real and fake samples. The second one is a pre-trained VGG-19 network and takes care of preserving the stylistic similarity with the original texture. The work proves a possibility to generate higher resolution, high-quality textures with preserved structures.

In addition to high-quality textures, generating related segmentation maps for dataset augmentation is necessary. The MaterialGAN (Guo et al., 2020) introduces a GANs modification for generating realistic SVBRDF parameter maps. They used the dataset from Deschaintre (Deschaintre et al., 2018) for training. It contains 155 SVBRDFs with a high resolution of 4096×4096 pixels. It was then augmented by blending multiple SVBRDFs to generate 256×256 resolution patches at a random position, rotation, and scaling. The generated result is 9-channelled with 3 channels for a fraction of incident light reflected from the surface, 2 for the surface orientation of the geo-

metric object, 1 for the roughness, and 3 for a fraction of incident light reflected from the surface. They used StyleGAN 2 as the baseline architecture.

Specialized and modified GAN architectures can have enough learning capacity to generate histological data. A PathologyGAN (Quiros et al., 2019) focuses on generating realistic histological images. The variability of data is introduced from two different training datasets, H&E colorectal cancer tissue from the National Cancer Center (NCT, Germany), H&E breast cancer tissue from the Netherlands Cancer Institute (NKI, Netherlands), and Vancouver General Hospital (VGH, Canada). In total, it contains 86 whole slide images and 576 tissue microarrays. They used BigGAN as the underlying architecture, which they augmented with a mapping network from StyleGAN, a style mixing regularization, and a relativistic mean as a loss function for the discriminator.

StyleGAN is also used for prostate cancer data synthesis (Daroach et al., 2022). However, the main focus is on the trained latent space of the StyleGAN to label the PCa regions according to the pathologist annotations. The pathologist attached a label to each of the model-generated realistically-looking patches. These labels then defined the regions in the original latent space from which sampled noise-generated histology images were always of the latent-space class. Therefore the StyleGAN-based solution is able to synthesize sample patches of specified prostate cancer classes. However, they still required help from a pathologist to annotate generated patches without further medical information about the sample, which may have introduced an error.

This paper presents a StyleGAN-based solution for selectively synthesizing epithelial cells, lymphocytes, macrophages, and neutrophils in the lungs, prostate, kidney, and breast. The result of our model is an RGB image with a segmentation map of cells' pixel positions and classes.

3 METHOD

The main goal of our method is to generate quality histology images with associated cell multi-class segmentation masks. GAN is the current, massively applied deep learning architecture framework suitable for this problem. According to the related work, GANs can generate non-stationary textures, medically valid histological data, related maps, and annotated segmentation masks. We based our generator architecture on StyleGAN.

Initial data are necessary to train the generator, so we chose the MoNuSAC dataset (Verma et al., 2020).

It contains TMA images with their annotated segmentation masks. The dataset consists of 4 cell types responsible for diagnosing stages and severity of lung, prostate, breast, and kidney cancer. Each segmentation mask contains information about the classes of cells and the organ. We use their initial color classes of the cells: red, yellow, green, and blue for epithelial, lymphocytes, macrophages, and neutrophils.

To validate the results and investigate the influence of the synthetic data used on training for segmentation, we employed the standard segmentation network UNet.

3.1 Generative Model

The tissue visuals depend on the organ, so the synthesis method must preserve its tissue characteristics. The standard input for the GAN network is sampled Gaussian noise. Therefore, we extended the StyleGAN architecture with an idea from Auxiliary Classifier GAN (ACGAN) (Odena et al., 2017). The organ class is global information we represent by a one-hot encoded vector, which sets the generator for the intended organ visual. The cell classes are specific to the location in the tissue. We do not pre-set the segmentation mask defining the location of cells. We use only the one-hot encoded vector to specify the expected classes the model should generate. To force the generator to synthesize only specified classes is the job of the discriminator. Also, to preserve the input information about classes, we modify the ACGAN approach and add the cell and organ information to every $2n$ layer of the StyleGAN mapping network as is shown in Figure 1. The result of the mapping network is the style vector used in adaptive instance normalization in generator layers. The generator architecture, random noise vector, constant vector of ones, and blending alpha values for progressive growing are the same as in the original StyleGAN paper.

The generator's output and the discriminator's input is an image with 6 channels (2x RGB). The first three channels present the generated histological image, and the last three channels present the generated segmentation mask of that image using the dataset's predefined colors per class. We need to modify the discriminator to force the generator to train according to the input class information. The standard regression is to distinguish real and fake images. Improved Wasserstein loss is applied to reduce the chances of Mode Collapse. The discriminator now requires also two additional classifiers. One classifier classifies the organ type - the class of the whole tissue segment, which is activated by softmax on the output layer and trained against multi-class categorical cross-entropy.

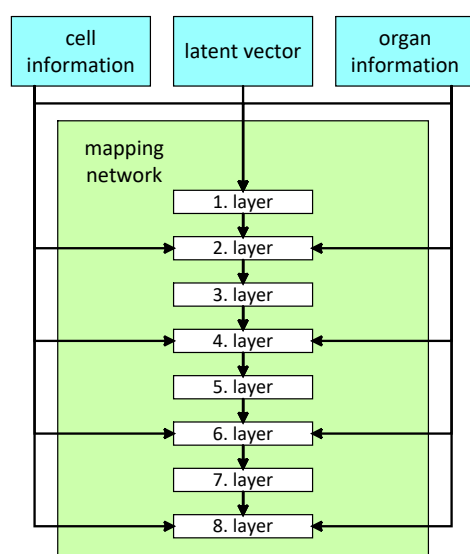


Figure 1: A modified architecture of StyleGAN mapping network with ACGAN class information input values.

The second classifier determines the classes of generated cells with segmentation masks. One image can contain cells of multiple classes, so we use sigmoid activation and multi-label binary cross entropy loss function. These values are compared to the input one-hot vectors. The architecture of a whole generative adversarial framework is in Figure 2.

The generated annotation masks contain a certain amount of noise, cell borders are sometimes unclear, and there can be multiple labels on the area of the same cell, even though the cell visual clearly represent only one class. We use simple post-processing by color normalization and morphological transformation to increase the quality of segmentation masks. The first step is color normalization to unify color segments into single-class clusters. Second, morphological closing enlarges areas, fills holes in cells, and removes unwanted details. Figure 3 visualizes the effect of post-processing of an example mask. Finally, we transform the pixel color values into 4 classes of regression for the computation of a loss function. A black background is 0, red is 1, green is 2, blue is 3, and yellow is encoded as 4.

4 EVALUATION

We evaluate both quantitative and qualitative results of the generated histological images along with their annotated segmentation masks. The impact of the synthetic data generated by our GAN is discussed over the segmentation results of the trained segmentation model.

4.1 Synthesis of Histological Data and Annotated Segmentation Masks

The presented images are the results of a model that took 192 hours (8 days) to train. The time distribution over the training of individual resolutions is the following: from 4 to 64 pixels took 24 hours, up to 128 pixels took 48 hours, and up to 256 pixels took 120 hours. Throughout the training, the values of all loss functions were balanced, and we did not observe any significant fluctuations. We expect that the longer training with further progressive upsampling would increase the quality of tissue and individual cells.

For the quantitative evaluation of our generative model, we use standard metrics Frechét Inception Distance (FID), Inception Score (IS), and Kernel Inception Distance (KID). Accuracy and Recall measure the classification score of the discriminator. All quantitative metrics are displayed in Table 1. The PathologyGAN (Quiros et al., 2019) achieved an FID of 32.05 on a different histology dataset to compare the results to related work.

The qualitative evaluation took place in the presence of a pathologist. After several moments they could determine that some images did not look realistic, but others could not distinguish from real samples. In some cases, the generated cells' structure looked similar to the real samples. The sampled cells contained the nucleus and preserved cytoplasm, and the structural placement of cells also looked realistic. However, to classify them, the pathologist stated, they would require the original tissue sample to see the whole tissue's structure. Therefore, for the dataset augmentation, we consider the results satisfactory. Some examples are documented in Figure 4. The post-processing of the segmentation maps helps to improve the quality and cell-border accuracy and to generate more exact cell types. The input information about cell classes and organs provides strict constraints for the generator.

4.2 Dataset Augmentation

We evaluated the impact of our generated synthetic data on the performance of a segmentation network. We use the MoNuSAC dataset and synthetic data generated by our trained generator. This section investigates and compares the effect of the synthetic data on the training and the segmentation predictions of UNet model evaluated on the MoNuSAC test subset.

We use the focal loss function, which is often used for training multi-label segmentation, to investigate the effect of synthetic data on the training of segmentation and prediction of multi-label masks.

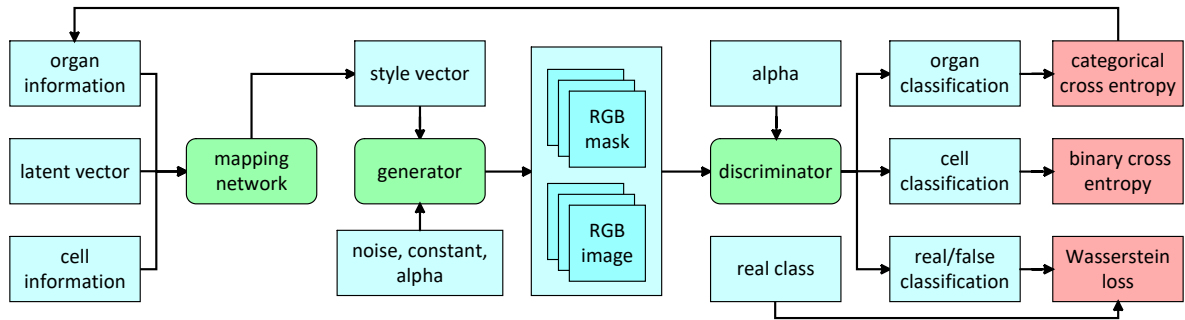


Figure 2: The architecture of used GAN framework. Generator inspired by StyleGAN. Discriminator modified to be auxiliary classifier.

Table 1: Metrics for synthetic data and classification performance of discriminator. The low values of the discriminator do not degrade the quality of the generated images.

Synthesis	IS	FID	KID	Accuracy	Recall
Images	4.126 ± 0.185	84.973	0.049 ± 0.001	0.254	0.365
Masks	2.711 ± 0.067	51.211	0.034 ± 0.001	0.867	0.484

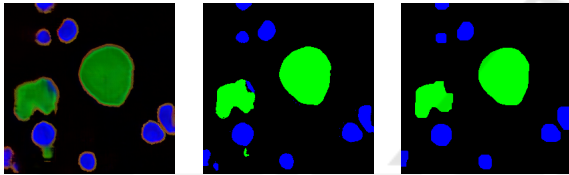


Figure 3: Post-processing of segmentation mask.

We train 3 models with the same architecture on batches with different real and synthetic data amounts. The datasets distributions of the 3 experiments are following:

1. Vanilla baseline model with only real samples — 1 : 0 real:synthetic
2. A sample of synthetic data that preserves a majority of real samples — 3 : 1 real:synthetic
3. Balanced dataset with the same distribution of real synthetic data — 1 : 1 real:synthetic.

The training dataset contains 3356 samples. 1656 are from the MoNuSAC dataset, and 1700 images are synthetic. To preserve the number of total training samples, the added amount of synthetic images is compensated by removing the same amount of random real samples. The test dataset contains 414 images, and all of them are from the MoNuSAC dataset. To preserve the same training conditions, the training hyperparameters of every model were the same, so the performance difference is affected only by the synthetic data.

- number of epochs: 20,
- batch size: 32,
- number of steps: 51,

- optimizer: Adam,
- learning rate: 0.001.

The courses of tracked training are displayed in Figure 5 and 6. Based on the results, we conclude that the model with 1 : 1 equal distributions had the biggest problems during training, leaving with the highest loss and the lowest accuracy. The model with a 3 : 1 smaller sample of synthetic data performed similarly to the vanilla model trained only on real data.

The qualitative results of the models are displayed in Figure 7. The segmentation masks produced by the model with a small amount of synthetic data achieved the best performance and even overcame the precision of the model trained only on real samples. Unexpectedly the model with an equal distribution suffers from over-segmentation, which makes it too inaccurate compared to the other two.

The quantitative metrics, Intersection over Union (IoU), Dice, and Hausdorff distance, are in Table 2. The quantitative performance measurements confirm the qualitative expectations that the model with an equally distributed dataset performed the worst. The model with an augmented dataset achieved similar and slightly better results than the vanilla model.

Table 2: Classification score depending on real to fake images rate in the training dataset.

UNET	IoU	Dice	Hausdorff
model data 1 : 0	0.556	0.815	16.455
model data 3 : 1	0.577	0.817	14.736
model data 1 : 1	0.459	0.704	17.921

To evaluate the performance of the segmentation, we analyzed the specific cell classes. Table 3 demon-

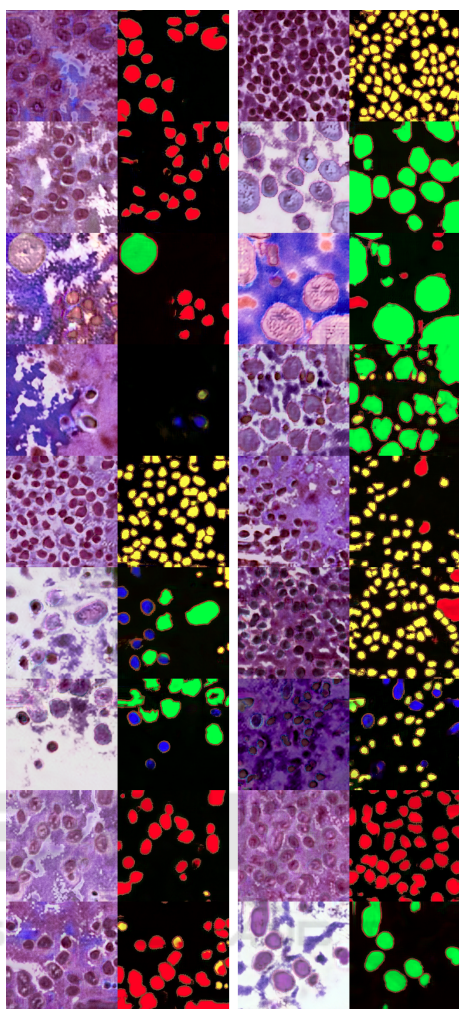


Figure 4: Direct results generated by our network without the post-processing.

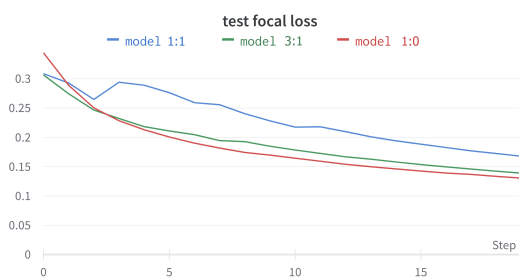


Figure 5: Progress of focal loss on a validation/test set during the training.

strates that in individual cases, the model trained on the dataset with a small amount of synthetic data even achieved better segmentation than on only real data. This could have happened by adding new data with higher variability of shapes and structures than the original dataset.

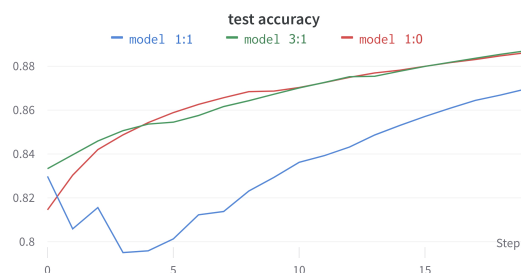


Figure 6: Progress of accuracy on a validation/test set during the training.

Table 3: Dice scores for each cell class and background.

Classes	Model 1 : 1	Model 3 : 1	Model 1 : 0
Background	0.820	0.861	0.877
Epithelial	0.401	0.577	0.570
Lymphocytes	0.203	0.362	0.360
Macrophages	0.418	0.514	0.406
Neutrophils	0.453	0.573	0.568

5 DISCUSSION

The proposed approach is tested and developed directly for histology data generation but can be used in different segmentation application domains with hierarchical data classes. Semantic cell segmentation allows us to explore the approach’s benefits and weaknesses and can be forgiving for some irregularities.

Compared to the related approaches like TilGAN (Saha et al., 2021) and PathologyGAN (Quiros et al., 2019), our approach can generate precise segmentation masks of several cell classes in different organs. A similar intermediate output can be found in the Un-supervised training GANs for segmentation in (Gardemayr et al., 2018), where the cycle GAN generates segmentation masks of circular or ellipsoid regions. Our approach uses supervised training, so the generated cells have various shapes depending on the annotated training data. The segmentation masks cell shape quality is improved by post-processing, however, at times, there can be cells with multi-label annotations, especially when the cell regions are bigger. This is difficult to correct because more cells can overlap at the same position since the tissue slice is a 3D volume. Also, the overall quality may be inferior to the PathologyGAN, but their model has BigGAN architecture which contains far more trainable parameters. Our model was trained using PyTorch (v1.11) on a desktop computer with NVIDIA RTX 3090, and the training took 24h for 4-64p, 33h for 128p, and 38h for 256p, summing up to 95h total training time.

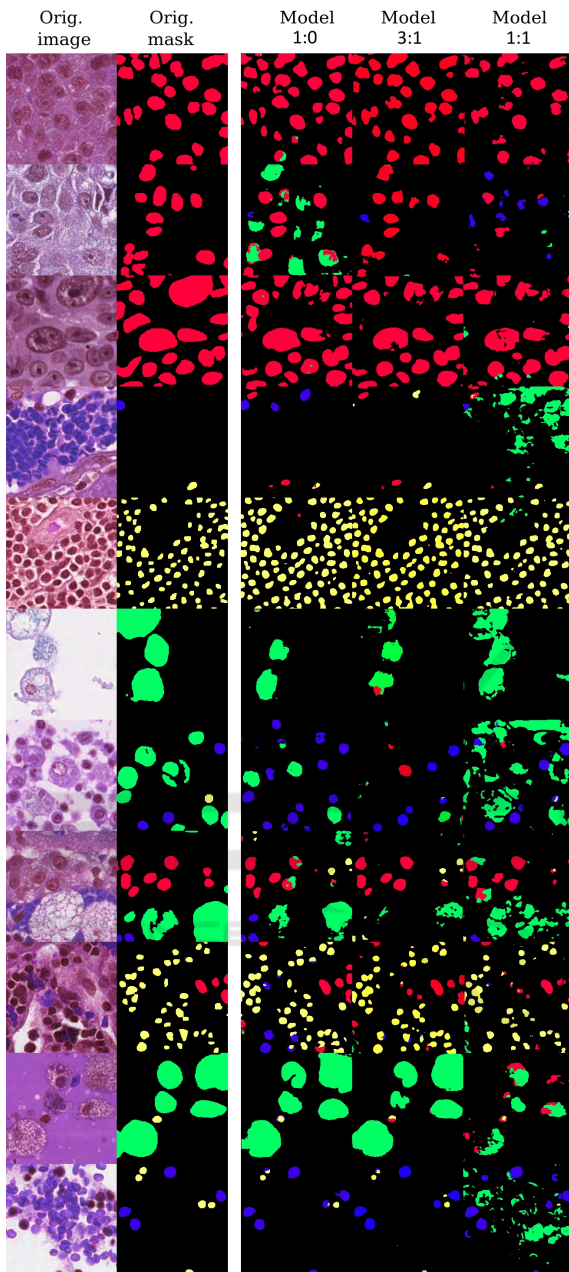


Figure 7: 2 left columns: Real images and ground truth annotations. 3 right columns: Segmentation predictions generated by our trained UNET models.

Compared to the state-of-the-art approaches, our StyleGAN modification with ACGAN allows us to control the generation of desired organ classes and cell types. Unfortunately, it is impossible to define the exact location of generated cells. However, we do not see this as a disadvantage because the generated images have sufficient variability, and to augment the dataset, it is not important to have total control over cell positions, and it is easier to leave it unat-

tended. Additionally, without explicit control of cell positions, the network can learn if there is a specific spatial distribution of cells in the real data.

The current maximal size of generated tissue patch is 256p, which is briefly large enough for most of the segmentation networks. It is also possible to increase the size of a generated patch by continuing training with progressive growing or to merge the generated patches by tiling.

The presented model was trained on the MoNuSAC dataset with 36000 hand-annotated cells, which is a large amount of manually made annotations. However, we expect the generator to learn the cell structures, tissue texture, and segmentation masks even if they would be annotated semi-automatically from, for example, QuPath (Bankhead et al., 2017).

6 CONCLUSION

This study presented a novel approach to histological datasets augmentation by generating images with corresponding annotations. The qualitative evaluation by the pathologist concluded that the images look similar to real tissue microarrays. Even though the synthetic data may have unrealistic artefacts, the data can be used to augment training datasets. We evaluated the impact on training through several experiments and three trainings where we observed the progress of the Focal loss function and classification accuracy of cell classes. The experiments proved that even a small amount of synthetic data might improve the final performance of a model. Also, the excessive amount of synthetic data can add bias to the dataset and hurt the generalization of the model.

In conclusion, we modified the StyleGAN architecture of auxiliary classification, so it is possible to control the generated cell type and organ type. The modification required adding two input layers to the mapping network and two classifiers to the discriminator. In order to better preserve the information about the cell and organ type, we appended the input information to every $2n$ layer of the mapping network.

We used MoNuSAC dataset and our synthetic generated data to evaluate the impact of the data augmentation on the training of the segmentation model. We trained three models with different real and synthetic data distributions. We set constant hyperparameters for each training to maintain objectivity. The model with a small amount of synthetic data achieved better results than vanilla training on only real data.

Future studies should consider increasing synthesized images' quality and improving the model and

output resolution. It could be beneficial for the pathologist to see the whole tissue structure, not only the detail of some selected cells.

ACKNOWLEDGEMENTS

This work was partially supported by STU Grant Scheme for support excellent teams of young researchers and Cooperation (Financial support) with Siemens Healthineers Slovakia.

REFERENCES

- Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., et al. (2017). Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7.
- Beers, A., Brown, J., Chang, K., Campbell, J. P., Ostmo, S., Chiang, M. F., and Kalpathy-Cramer, J. (2018). High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv preprint arXiv:1805.03144*.
- Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.
- Daroach, G. B., Duenweg, S. R., Brehler, M., Lowman, A. K., Iczkowski, K. A., Jacobsohn, K. M., Yoder, J. A., and LaViolette, P. S. (2022). Prostate cancer histology synthesis using stylegan latent space annotation. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 398–408, Cham. Springer Nature Switzerland.
- Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., and Bousseau, A. (2018). Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15.
- Gadermayr, M., Gupta, L., Klinkhammer, B. M., Boor, P., and Merhof, D. (2018). Unsupervisedly training gans for segmenting digital pathology with automatically generated annotations. *ArXiv*, abs/1805.10059.
- Guo, Y., Smith, C., Hašan, M., Sunkavalli, K., and Zhao, S. (2020). Materialgan: reflectance capture using a generative svbrdf model. *arXiv preprint arXiv:2010.00114*.
- Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.
- Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR.
- Quiros, A. C., Murray-Smith, R., and Yuan, K. (2019). Pathologygan: Learning deep representations of cancer tissue. *arXiv preprint arXiv:1907.02644*.
- Saha, M., Guo, X., and Sharma, A. (2021). Tilgan: Gan for facilitating tumor-infiltrating lymphocyte pathology image synthesis with improved image classification. *IEEE access: practical innovations, open solutions*, 9:79829 – 79840.
- Štepec, D. and Skočaj, D. (2020). Image synthesis as a pretext for unsupervised histopathological diagnosis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 174–183. Springer.
- Verma, R., Kumar, N., Patil, A., Kurian, N. C., Rane, S., and Sethi, A. (2020). Multi-organ nuclei segmentation and classification challenge 2020. *IEEE transactions on medical imaging*, 39(1380-1391):8.
- Zhou, Y., Zhu, Z., Bai, X., Lischinski, D., Cohen-Or, D., and Huang, H. (2018). Non-stationary texture synthesis by adversarial expansion. *arXiv preprint arXiv:1805.04487*.