Diffusion models for nuclei segmentation in low data regimes

Konstantinos Alexis, Stergios Christodoulidis, Dimitrios Gunopulos, Maria Vakalopoulou

DIFFUSION MODELS FOR NUCLEI SEGMENTATION IN LOW DATA REGIMES

Konstantinos Alexis[1,3], *Stergios Christodoulidis*[2], *Dimitrios Gunopulos*[1], *Maria Vakalopoulou*[2]

[1] Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece
[2] CentraleSupélec, University Paris-Saclay, France and Archimedes/Athena RC, Greece
[3] Information Management Systems Institute, Athena Research Center, Greece

ABSTRACT

Nuclei detection and characterization in histopathological tissue assessment is of utmost importance for different clinical workflows, such as the characterization of tumor microenvironments. Utilizing robust computational models for such a task could allow streamlining the process. However, obtaining accurate segmentation maps for histopathological slides can be quite tedious and expensive, while also being subject to inter/intra-reader variability. Learning robust and precise segmentation models from only a small amount of data would be therefore a very interesting alternative to fully supervised methods relying on a huge amount of annotations. Inspired by diffusion models' recent advances, this paper proposes a method for obtaining nuclei segmentation maps under low data regimes. In particular, diffusion models are used for learning powerful pixel-level representations of digital pathology patches that could require only a few amounts of annotated data to provide multiclass segmentation maps of different nuclei. Various insights about the use of these models for the representation of digital pathology patches are provided. Comparisons with other self-supervised and fully supervised methods highlight the advantages of the use of these models for nuclei segmentation.

***Index Terms*—** multiclass nuclei segmentation, histopathology, denoising diffusion probabilistic models.

1. INTRODUCTION

Histopathology is one of the gold standards for the diagnosis and treatment selection for cancer patients. However, assessing these gigapixel-sized images is quite challenging and can suffer from inter-observer variability. Digital pathology seeks to develop automatic methods, aiming to simplify clinical practices and standardize treatment decisions across various centers and protocols. While these tools could improve workflows for clinicians, incorporating them into clinical practice isn't always straightforward. One example that highlights this need is the problem of nuclei quantification and segmentation on gigapixel-sized images, which is of utmost importance for the categorization of different types of cancers and the selection of different treatment protocols, yet it remains quite challenging.

Recently, deep learning architectures could provide robust models for a variety of applications, including different medical image analysis tasks [1]. Semantic segmentation of medical images is one of them, with recent methods based on U-Net-like [2] or Vision Transformer-based [3] architectures reporting impressive performance. However, these methods usually require a lot of pixel-wise annotations to be trained, something that is not always possible for medical datasets on which the availability of annotations can be quite hard to acquire and expensive. Indeed, this is one of the main problems that slow down the adaptation of deep learning models to clinical practice and reduce their robustness.

Different directions have been explored in the recent years to address the problem of deep learning training on low data regimes. Among the different approaches, methods based on few-shot learning [4], using only a small number of examples enable the model to generalize well to new, unseen data despite the limited amount of training samples. Moreover, methods based on self-supervision and contrastive learning are currently gathering the attention of the community, providing interesting alternatives for the training of powerful feature representations that could be adapted to different tasks, including image segmentation. Additionally, denoising diffusion probabilistic models (DDPM) [5] could also be used to provide useful representations, something that has only recently started to be explored by the community [6].

In this work, we investigate whether diffusion models are able to learn semantically meaningful representations in the context of digital pathology. Our approach follows the intuition that the unsupervised training of generative diffusion models results in robust features, which can be useful for a number of downstream tasks such as semantic segmentation, especially in limited-annotation schemes. We focus on the DDPM class of models, and we re-purpose them for a non-generative task, aiming to assess their performance on a discriminative task for the histopathology image domain. The main contributions of this work are summarized as follows:

- We train a DDPM for representation learning of digital

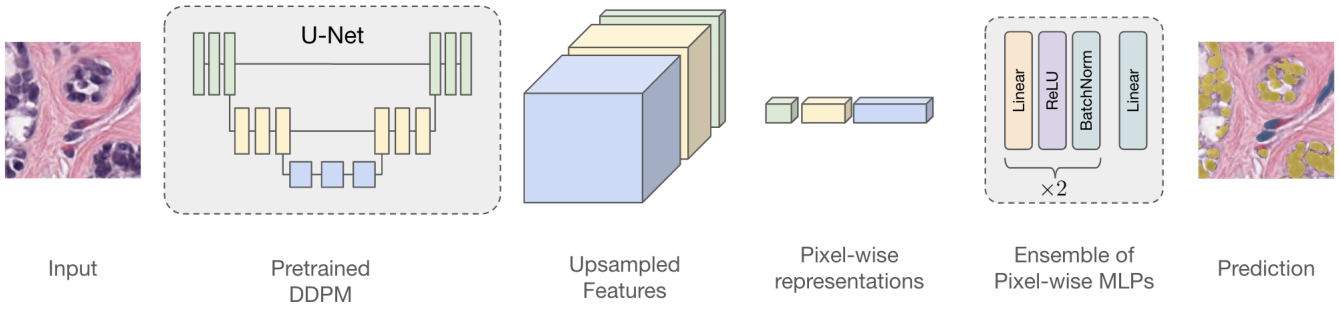

Fig. 1. Overview of the proposed method. The first step of the method trains in an unsupervised way a DDPM using a pancancer dataset of digital pathology patches. Then, the features of the different levels of the pretrained DDPM are upsampled, and pixel-wise representations are obtained. These pixel-wise features are then classified in each of the available classes with an ensemble of MLPs.

pathology slides. Our model is trained in an unsupervised way using patches extracted from different cancer types.
- We demonstrate that DDPMs could provide better representations for semantic segmentation tasks compared to other self-supervised methods.
- We provide insights about the use of DDPMs in the context of digital pathology and nuclei segmentation, where the structures are very small, making self-supervised segmentation methods particularly challenging.

2. RELATED WORK

Automatic nuclei segmentation in digital pathology has gathered the attention of the community due to a wide range of clinical applications. Indeed, a single tissue slide typically contains around a million nuclei, making their quantification and annotation very tedious for pathologists. Various publicly available datasets have already been proposed to ease this need, providing a very interesting playground for deep learning algorithms. CoNSeP [7], MoNuSeg [8] and PanNuke [9] are only some of the publicly available datasets that provide pixel-wise nuclei segmentation. However, even though these datasets are available, trained models usually fail to generalize well on other samples due to big stain variability between centers and protocols. Indeed, the generalization of these algorithms is quite challenging [10], highlighting the need for methods that could be easily trained with only a few amount of annotated data.

The most prevalent nuclei segmentation methodologies currently depend on manually acquired, careful pixel-based annotations of nuclei [7, 11] tailored to specific staining protocols. Some semi-supervised approaches, like [12], have been proposed to address this requirement, yet they need manual interactions, making their application at the whole slide level time-consuming. On the other hand, existing unsupervised segmentation methods, such as those utilizing color clustering, exhibit inadequate performance, precluding their use in clinical settings. Researchers have also explored self-supervised learning for nuclei segmentation. In [13], the authors train a network to precisely classify the magnification of an input tile using an attention module and demonstrate that the attention maps can be employed to generate detection maps of nuclei in H&E staining, subsequently transformable into nuclei segmentation maps. Moreover, in [14], the authors proposed a dense contrastive scheme for the segmentation of nuclei. Even if, self-supervised methods could provide very interesting research directions, most of the time are not able to provide multiclass segmentation maps.

In the context of the limited amount of available data, powerful pretrained models such as DINOv2 [15] and other recent foundation models would also be used to extract powerful representations of the data and provide segmentation maps as a downstream task. In a similar idea, DDPMs has been recently introduced for the development of generative representation learning [16, 6]. Nonetheless, even if these models had been proposed and validated for natural images, their application on medical data and, in particular, digital pathology images is not assured. Especially, in the case of digital pathology and nuclei segmentation and identification, the very small structure of the objects of interest makes the use of such models very challenging. In this work, we investigate exactly this point, providing new insights for the use of DDPM in digital pathology.

3. METHODOLOGY

DDPM Background. In this work, we adopt the formulation of Denoising Diffusion Probabilistic Models as presented in [5]. DDPMs are able to transform noise into data samples by learning to reverse a progressive forward noising process, which can be described as:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \qquad (1)$$

Making use of Gaussian parametrization allows directly getting an arbitrary timestep's x_t through:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The reverse diffusion process, commonly parameterized through a U-Net network variant $\epsilon_\theta(x_t, t)$ is trained to predict the noise added at each timestep, substantially implementing a multi-step denoising task. Once trained, this model can be used to generate realistic data samples just from random noise by solving a reverse process:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

DDPM Representations. Our method relies on a trained DDPM on histopathology data using the approach proposed in [18]. By training an in-domain DDPM, we aim to build a robust encoder from which we will be extracting image representations. We note, that this training procedure is unsupervised; no manual annotations are needed, thus making the approach especially useful for low-data settings. After training such a model, we employ it to extract representations in order to investigate whether they can capture meaningful information related to nuclei. For this task, our method is based on [6], where we corrupt input images with noise corresponding to selected timesteps and extract the feature maps that the hidden blocks of the U-Net decoder output, when we pass the corrupted images through them. We choose the timesteps and blocks to keep in our setting by combining the intuition provided by [6] with our experimental results. Specifically, the most descriptive features correspond to the latter timesteps of the reverse diffusion process, while regarding the blocks, the intermediate ones are found to perform best.

Semantic Segmentation. To assess the information captured by these representations, we use them for a downstream task, namely, multiclass semantic segmentation of nuclei in histopathology images. An overview of the method is presented in Fig. 1. We follow the approach of [6], firstly interpolating the extracted hidden activations for an image into the spatial dimensions of the initial input and then stacking them into a single feature map. Thus, each pixel is finally represented by a feature vector of size equal to that of the whole feature map along its concatenated axis. Under this formulation, the encoded pixels are then passed to an ensemble of MLP classifiers, where the predicted class for the pixel is determined via a majority voting. Despite its simplicity, this segmentation approach is quite robust, providing valuable insights into the knowledge embedded in the proposed pixel-level features.

4. IMPLEMENTATION DETAILS

Dataset. We conducted our experiments using the PanNuke [9] dataset. Initially, we merged the three independent folds comprising the entire dataset, totaling 7901 images, into a single dataset. The intuition behind this decision is to help provide a sufficiently large dataset split for pretraining the DDPM within the specific data domain. Thus, we then split 80% of the dataset (6321 patches) for pretraining the DDPM while retaining the rest 20% (1580 patches) for evaluation on the semantic segmentation task to avoid data leakage. This was done by splitting these 1580 patches into 448 and 1132 for training and testing, respectively. Every time we split the dataset, we maintain the nuclei distribution of the initial dataset, keeping representative samples from all nuclei types.

Training. We pretrain our DDPM using a diffusion process of 1000 steps with a linear noise schedule and a learning rate of $1e-4$. After assessing multiple model checkpoints based on the training steps, we determined that the model checkpoint at 50K steps performed best in downstream segmentation. Notably, while the 100K model checkpoint showed better sampling performance i.e., generating more realistic samples, it did not lead to improved segmentation performance. Taking into consideration the specific characteristics of our dataset i.e., nucleis' small structure, variability and heterogeneity, we finetune over the selection of timesteps and blocks (comprising hyperparameters of the method) in order to come up with an optimal configuration for our task. Thus, we run all our experiments extracting representations for timesteps $[50, 150, 250]$ and from $[6, 8, 10, 12]$ blocks of the DDPM, resulting in pixel-level features with a shape of 2688. Extracting features from earlier timesteps or even maintaining a broader range of blocks seemed to degrade the segmentation performance. Moreover, to account for the background class covering a high percentage of the pixels among histopathology images, we train the MLP classifiers keeping only 10% of the background pixels, avoiding the background class signal from overwhelming the training process. Finally, we train and test our segmentation models on 448 and 1132 images, respectively.

5. EXPERIMENTAL RESULTS

Baselines. We compare the performance of our method against a series of baselines for the task of semantic segmentation. U-Net refers to the vanilla model, while Attention U-Net [19] incorporates attention mechanisms, to focus on informative regions by dynamically weighting feature maps during the encoding and decoding phases. The DINOv2 variants denote Vision Transformer [20] models pretrained on a vast dataset through self-supervised learning. The Linear and SETR-PUP variants refer to the segmentation layers on top of the pretrained backbone model. The former is described in [15], while the latter integrates multiple deconvolution layers as a decoder to generate segmentation outputs [21]. For DINOv2 models we train only the task layers; the rest of the models parameters are kept frozen. $\text{DDPM}_{\text{ImageNet}}$ follows the same approach as our proposed method with the exception of being pretrained on ImageNet [22].

Method	Neoplastic	Inflammatory	Connective	Dead	Epithelial	Mean
Fully Supervised methods						
U-Net [17]	0.6975	0.5151	0.486	0.2363	0.6145	0.5099
Attention U-Net [17]	0.7039	0.5351	0.4833	0.2584	0.6445	0.52504
Pretrained Models with low data regimes						
$DINOv2_{Linear}$ [15]	0.5511	0.38	0.321	0.0298	0.4339	0.3432
$DINOv2_{SETR\text{-}PUP}$[15]	0.6368	0.5094	0.4284	0.1527	0.5341	0.4523
$DDPM_{ImageNet}$	0.5445	0.426	0.3344	0.1016	0.39	0.3593
$DDPM_{PanNuke}$ (Ours)	0.6211	0.4945	0.4094	0.1919	0.5601	0.4554

Table 1. Quantitative results for the different types of nuclei and the mean performance. The table presents the intersection over union (IoU) results on the test set. Methods are grouped as supervised, semi-supervised, and diffusion-based.

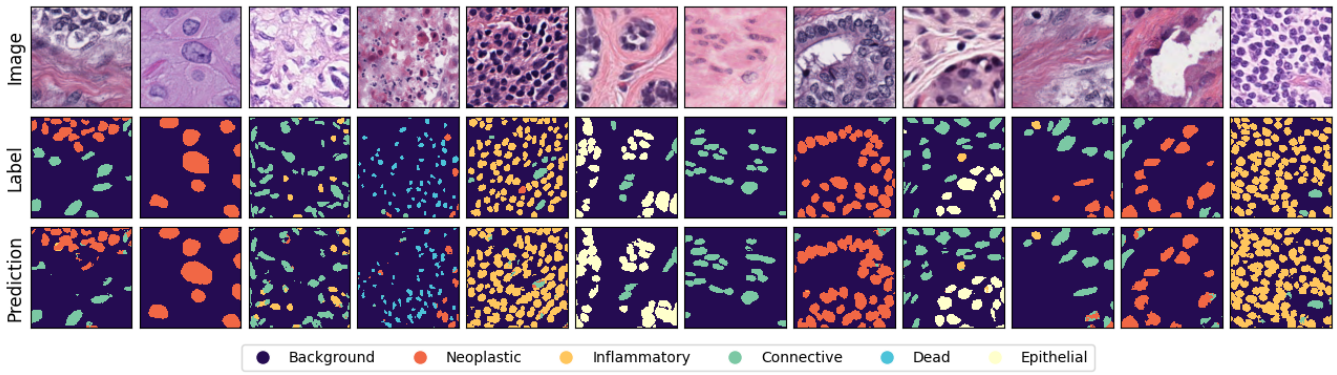

Fig. 2. Qualitative results of the proposed DDPM method for different types of nuclei. The first row depict different patches, the second the ground truth with different colors, and the last one our prediction.

The segmentation results are presented in Table 1. While it is clear that fully supervised methods yield superior models, their dependence on abundant data might pose limitations in various scenarios. Notably, within low data regimes, DDPMs exhibit a slight performance edge over DINOv2 counterparts, despite the latter being pretrained on a significantly larger dataset. Furthermore, when comparing the two DDPMs, the importance of in-domain pretraining for achieving enhanced downstream performance is underscored, particularly in the segmentation of very small structures such as different classes of nuclei.

Proposed DDPM	
# of Images	Mean IoU
28	0.3298
56	0.3667
112	0.3964
224	0.4341
448	0.4554

Table 2. Ablation on the number of training images used for the segmentation task.

Table 2 demonstrates an ablation of our model, indicating how the performance of the proposed DDPM method consistently improves with an increase in the number of training images, with the best results achieved when utilizing all 448 training images. Qualitative evaluation of this selected model is presented in Fig. 2 demonstrating different histopathology patches from the test set (first row) alongside their ground truth segmentation (second row) and our predictions (third row). Overall, our model demonstrates consistent and precise performance across all the different patches originating from different cancers and locations.

6. CONCLUSIONS

We investigate the representations learned by DDPMs in the histopathology domain, affirming their ability to capture semantically meaningful information, as demonstrated in the segmentation setting. The proposed method leverages the generative pretraining of DDPMs, producing robust representations without requiring a large number of manual annotations, thereby enhancing their generalizability and applicability in scenarios with limited access to labels. Our results move towards the direction of using generative models for discriminative tasks. Future works include the further testing of our method on different segmentation tasks of digital pathology and the extension of the method integrating spatial relations between pixels to refine the segmentation map.

7. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required. All datasets used in this study are publicly available.

8. ACKNOWLEDGMENTS

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program and the EU's Horizon Programme call, under Grant Agreement No. 101093164 (ExtremeXP). D. Gunopulos has been partially supported by the CoDiet project which is funded by the European Union under Horizon Europe grant number 101084642 and supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee.

9. REFERENCES

[1] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, 2022.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*. Springer, 2015.

[3] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Engineering Applications of Artificial Intelligence*, vol. 126, 2023.

[4] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, 2020.

[6] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *ICLR*, 2022.

[7] S. Graham, Q. D. Vu, S. E A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, 2019.

[8] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, PA Heng, J. Li, Z. Hu, et al., "A multi-organ nucleus segmentation challenge," *IEEE transactions on medical imaging*, vol. 39, no. 5, 2019.

[9] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," *arXiv preprint arXiv:2003.10778*, 2020.

[10] D. Tellez, G. Litjens, P. Bándi, W. Bulten, JM Bokhorst, F. Ciompi, and J. Van Der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical image analysis*, vol. 58, 2019.

[11] I. Kiran, B. Raza, A. Ijaz, and M. A Khan, "Denseres-unet: Segmentation of overlapped/clustered nuclei from multi organ histopathology images," *Computers in Biology and Medicine*, vol. 143, 2022.

[12] N.A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot, "Nuclick: a deep learning framework for interactive segmentation of microscopic images," *Medical Image Analysis*, vol. 65, 2020.

[13] M. Sahasrabudhe, S. Christodoulidis, R. Salgado, et al., "Self-supervised nuclei segmentation in histopathological images using attention," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

[14] J. Zhang, S. Kapse, K. Ma, P. Prasanna, M. Vakalopoulou, J. Saltz, and D. Samaras, "Precise location matching improves dense contrastive learning in digital pathology," in *International Conference on Information Processing in Medical Imaging*. Springer, 2023.

[15] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[16] K. Abstreiter, S. Mittal, S. Bauer, B. Schölkopf, and A. Mehrjou, "Diffusion-based representation learning," *arXiv preprint arXiv:2105.14257*, 2021.

[17] Z. Ye, B. Hu, H. Sui, M. Mei, L. Mei, and R. Zhou, "Dscanet: Double-stage codec attention network for automatic nuclear segmentation," *Biomedical Signal Processing and Control*, vol. 88, 2024.

[18] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, Curran Associates, Inc.

[19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[21] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021.

[22] J. Deng, W. Dong, R. Socher, LJ Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.