

Pre-Training with Diffusion models for Dental Radiography segmentation

Jérémy Rousseau, Christian Alaka*, Emma Covili*, Hippolyte Mayard*, Laura Misrachi*, and Willy Au

Allisone Technologies, Paris, France <https://www.allisone.ai/>
{jeremy, christian, emma, hippolyte, laura, willy}@allisone.ai

Abstract. Medical radiography segmentation, and specifically dental radiography, is highly limited by the cost of labeling which requires specific expertise and labor-intensive annotations. In this work, we propose a straightforward pre-training method for semantic segmentation leveraging Denoising Diffusion Probabilistic Models (DDPM), which have shown impressive results for generative modeling. Our straightforward approach achieves remarkable performance in terms of label efficiency and does not require architectural modifications between pre-training and downstream tasks. We propose to first pre-train a Unet by exploiting the DDPM training objective, and then fine-tune the resulting model on a segmentation task. Our experimental results on the segmentation of dental radiographs demonstrate that the proposed method is competitive with state-of-the-art pre-training methods.

Keywords: Diffusion · Label-Efficiency · Semantic Segmentation · Dataset Generation

1 Introduction

Accurate automatic semantic segmentation of radiographs is of high interest in the dental field as it has the potential to help practitioners identify anatomical and pathological elements more quickly and precisely. While deep learning methods show robust performances at segmentation tasks, they require a substantial amount of pixel-level annotations which is time-consuming and demands strong expertise in the medical field. Accordingly, many recent state-of-the-art methods [9,6,5,22,2,23] use self-supervised learning as a pre-training step to improve training and reduce labeling effort in computer vision.

Inspired by the renewed interest in denoising for generative modeling, we investigate denoising as a pre-training task for semantic segmentation. Denoising autoencoder is a classic concept in machine learning where a model learns to separate the original data from the noise, and implicitly learns the data distribution by doing so [16,17]. In particular, denoising objective can be easily defined pixel-wise, making it especially well suited for segmentation tasks [4].

* These authors contributed equally to this work

Recently, a new class of generative models, known as Denoising Diffusion Probabilistic Models (DDPM) [10,15,13], have shown impressive results for generative modeling. DDPM outperform other state-of-the-art generative models such as Generative Adversarial Networks (GANs) [8] in various tasks, including image synthesis [7].

DDPM learn to convert Gaussian noise to a target distribution via a sequence of iterative denoising steps, yielding impressive results in image synthesis outperforming GANs [7,8].

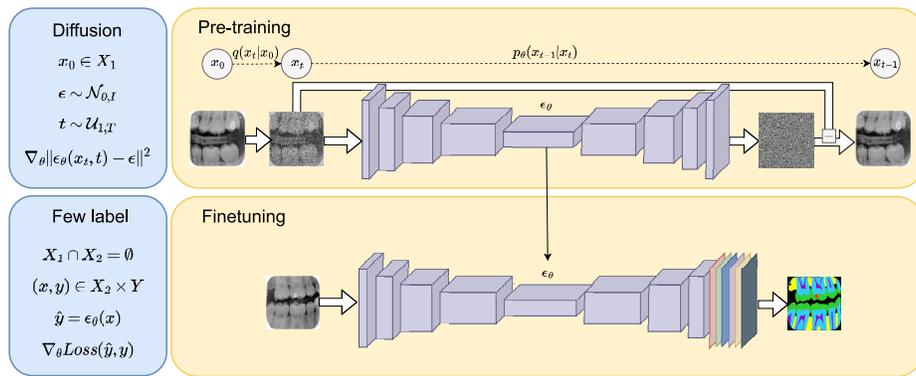


Fig. 1. PTDR method overview.

top - ϵ_θ is pre-trained on unlabeled dataset X_1 using the training procedure of DDPM [10]. *bottom* - ϵ_θ is then fine-tuned on a small labeled dataset X_2 . Y represents the set of ground truth semantic maps.

Following the success of DDPM for generative modeling, [1,18,19,20] explore their ability to directly generate semantic maps in an iterative process by conditioning each denoising steps with a raw image prior. [3] shows that DDPM are effective representation learners whose feature maps can be used for semantic segmentation, beating previous pre-training methods in a few label regime.

In this paper, we propose Pre-Training with Diffusion models for Dental Radiography segmentation (PTDR). The method consists in pre-training a Unet [14] in a self-supervised manner by exploiting the DDPM training objective, and then fine-tuning the resulting model on a semantic segmentation task.

To sum up our contributions, our method is most similar to [3] but does not require fine-tuning a different model after pre-training. The whole Unet architecture is pre-trained in one step at the difference of [4] which requires two. At inference, only one forward pass is used, making it easier to use than [3,1]. Finally, we show that our proposed method surpasses other state-of-the-art pre-training methods especially when only few annotated samples are available.

2 Methodology

2.1 Background

Inspired by Langevin dynamics, DDPM [10] formalize the generation task as a denoising problem where an image is gradually corrupted for T steps and then reconstructed through a learned reverse process. Generation is done by applying the reverse process to pure random noise.

Starting from an image \mathbf{x}_0 , the forward diffusion process iteratively produces noisy versions of the image $\{\mathbf{x}_t\}_{t=1}^T$, and is defined as a Gaussian Markov chain where $\{\beta_t \in (0, 1)\}_{t=1}^T$ is the variance schedule:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right) \quad (1)$$

A noisy image \mathbf{x}_t is obtained at any timestep \mathbf{t} from the original image \mathbf{x}_0 with the following closed form, let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ we have:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}\right) \quad (2)$$

When the diffusion steps are small enough, the reverse process can also be modeled as a Gaussian Markov chain:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right) \quad (3)$$

where:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \quad \sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (4)$$

with ϵ_θ the neural network being optimized.

The training procedure is finally derived by optimizing the usual variational bound on the negative log-likelihood, and consists of randomly drawing samples $\epsilon \sim \mathcal{N}_{0, \mathbf{I}}$, $\mathbf{t} \sim \mathcal{U}_{1, T}$, $\mathbf{x}_0 \sim q(x_0)$ and taking a gradient step on

$$\nabla_\theta \left\| \epsilon_\theta \left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t \right) - \epsilon \right\|^2 \quad (5)$$

2.2 DDPM for semantic segmentation

The proposed method is based on two steps. First, a denoising model is pre-trained on a large set of unlabeled data following the procedure presented in section 2.1. Second, the model is fine-tuned for semantic segmentation on few annotated data of the same domain by minimizing the cross-entropy loss.

Our method is similar to [3] which leverages a pre-trained DDPM-based model as a feature extractor. Their method involves upsampling feature maps

from predetermined activation blocks - from several forward passes at different timesteps - to the target resolution and training an ensemble of pixel-wise classifiers on concatenated feature maps. [3] showed that semantic information carried by feature maps highly depends on the activation block and the diffusion timestep. The latter are thus important hyper-parameters that need to be tuned for each specific semantic task. This method originally introduced in [24] - in the context of GANs - is well-suited for generative models feature extraction but does not leverage the DDPM architecture as PTDR does.

Our approach, by simply re-using the DDPM-trained denoising model for the downstream task, does not need extra classifiers and does not depend on activation blocks hyper-parameter. Moreover, PTDR fine-tuning and inference phases only require one forward pass in which the timestep is fixed to a predetermined value. To that extent, the proposed method is simpler both in terms of training and inference.

3 Experiments and Results

3.1 Experimental Setup

In our experiments, a Unet*¹ based DDPM is trained on unlabeled radiographs, the Unet* is then fine-tuned on a multi-class semantic segmentation task as illustrated in figure 1. We experiment with regimes of 1, 2, 5 and 10 training samples and compare our results to other state-of-the-art self-supervised pre-training methods. We used a single NVIDIA T4 GPU for all our experiments.

Datasets: Our main experiment is done on dental bitewing radiographs collected from partner dentists, see figure 2. The pre-training dataset contains 2500 unlabeled radiographs. Additionally, 100 bitewing radiographs are fully annotated for 6 classes namely: *dentine*, *enamel*, *bone*, *pulp*, *other* and *background* as semantic maps; and is randomly split into 10 training, 5 validation, and 85 test samples. There is no intersection between the pre-training and fine-tuning dataset. For our experiments, we use random subsets of the train set of size 1, 2, 5 and 10 respectively. Images are resized to 256x256 resolution and normalized between -1 and 1.

Pre-training: The Unet* implemented in pytorch is trained with a batch size of 2 and follows the training procedure of [7] with 4000 diffusion steps T . We use the official pytorch implementation of [7]. The training was performed for 150k iterations and we saved the weights at iteration 10k, 50k, 100k and 150k for fine-tuning comparison.

Fine-tuning: The batch size is set at 2. We use a random affine augmentation strategy with the following parameters: rotation angle uniformly sampled from

¹ Unet* denotes the specific Unet architecture introduced in [7]

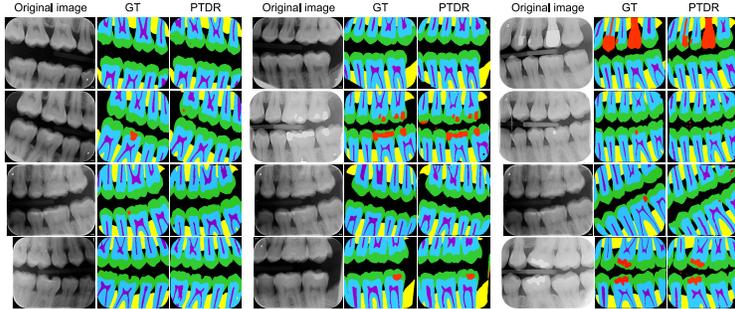


Fig. 2. Comparison on test dental bitewing radiographs of ground truth (GT) against predicted semantic maps from PTDR fine-tuned on 10 labeled images.

$[-180, 180]$, shear sampled from $[-5, 5]$, scale sampled from $[0.9, 1.1]$, and translate factor sampled from $[0.05, 0.05]$. Fine-tuning is done for 200 epochs using the Adam optimizer [11] with a learning rate of $1e^{-4}$, a weight decay of $1e^{-4}$, and a cosine scheduler.

Baseline methods: The DDPM training procedure is performed for 150k iterations and used for both PTDR and [3] which is referred to as DDPM-MLP for the next sections. We also pre-train a Unet* encoder with MoCo v2 [6] and then fine-tune the whole network on the downstream task. We refer to this method as MoCo v2. Finally, we pre-train a Swin Transformer [12] using SimMIM [22] and use it as an Upernet [21] backbone. We refer to this method as SimMIM. As the Swin backbone relies on batch normalization layers, we do not train SimMIM in the 1-shot regime. For all these methods, we use the same hyper-parameters as proposed in the original papers.

Evaluation metric: We use mean Intersection over Union (mIoU) as our evaluation metric to measure the performance of the downstream segmentation task.

3.2 Results

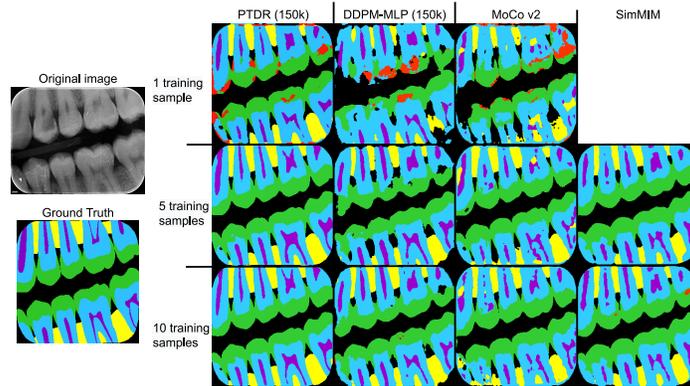
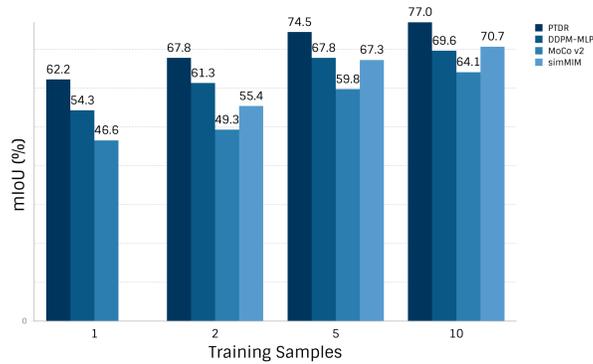
We compare our method with other baseline pre-training methods and compare their performances on the multi-class segmentation downstream task in the 10-labeled regime as shown in table 1.

Our method outperforms all other methods, improving upon the second-best method by 10.5%. Qualitative results on bitewing radiographs are shown in figure 3 with predicted semantic maps produced by all compared methods for 1, 5, and 10 training samples. For all regimes, predictions from our method are less coarse than others.

Label efficiency: In this experiment, we compare our method with baseline methods in different data regimes. Figure 4 illustrates the comparison between methods fine-tuned on 1, 2, 5, and 10 training samples.

Table 1. Comparison of pre-training methods when fine-tuned on 10 labeled samples

Model	Pre-training	mIoU
SwinUpperNet	–	59.58
	SimMIM [22]	70.69
Unet*	–	61.40
	MoCo v2 [6]	64.10
	DDPM-MLP [3]	69.64
	PTDR (ours)	76.96

**Fig. 3. Semantic maps** produced by different methods, PTDR, DDPM-MLP, MoCo v2 and SimMIM. The DDPM pre-training procedure is performed for 150k iterations. Semantic maps were produced by models trained on 1, 5, and 10 training samples to illustrate label efficiency.**Fig. 4. Label efficiency.** Comparison of pre-training methods when fine-tuning in several data (1, 2, 5, and 10 training samples).

Results show that our method yields better performance, in any regime, than all other pre-training methods benchmarked. On average, over all regimes, PTDR improves upon DDPM-MLP, its closest competitor, by 7.08%. Moreover, we can observe in figure 4, that our method trained on only 5 training samples outperforms all other methods trained on 10 samples.

Saturation effect: We explore the influence of the number of DDPM pre-training iterations on the per final segmentation performance. In figure 5, we observe strong benefits of pre-training between 10k and 50k iterations with an absolute mIoU increase of +7% for PTDR and +6% for DDPM-MLP. As we advance in iteration steps, the pre-training effectiveness decreases. For both methods, we observe that beyond 50k iterations, the performance *saturates* reaching a plateau. This suggests pre-training DDPM can be stopped before reaching ultra-realistic generative performance while still providing an efficient pre-trained model.

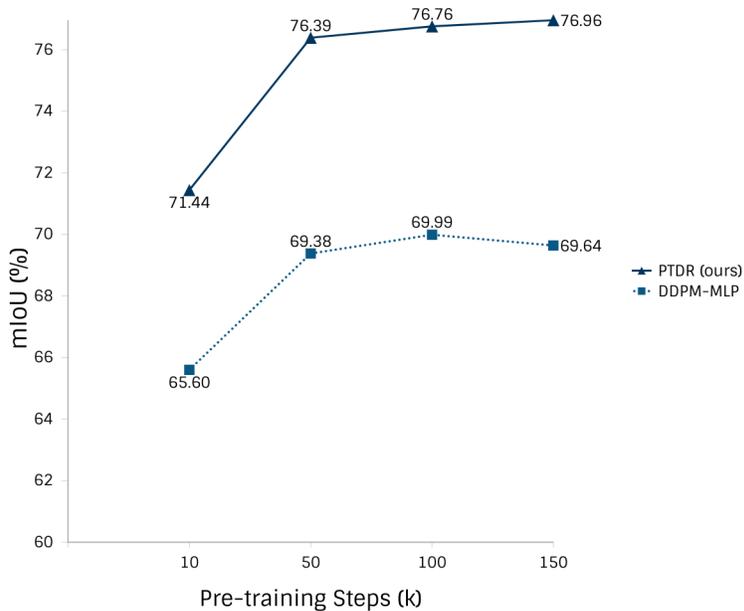


Fig. 5. Saturation effect. Impact of the number of pre-training steps on mIoU for PTDR and DDPM-MLP trained on 10 training samples.

Timestep influence: We investigate the influence of timestep, which conditions the Unet* and the amount of Gaussian noise added during the diffusion process. We empirically show in table 2 that timestep 1 is the optimal setup during fine-tuning. This is intuitive as this timestep corresponds to the first diffusion step during which images are almost not corrupted which mirrors the fine-tuning setup on raw images. We did not find any benefits from letting the network learn the timestep value. However, it is worth mentioning that when we do so, the timestep converges to 1.

Table 2. Influence of timestep value on PTDR’s fine-tuning performance

Timestep value	1	100	1000	2000	4000	learnt
mIoU	76.96	76.94	76.61	74.86	73.60	76.80

Generalization capacity: In appendix A, we further investigate the generalization capacity of our method to another medical dataset.

Dataset generation: In appendix B, we qualitatively illustrate the method’s ability to generate a high-quality artificial dataset with pixel-wise labels.

4 Conclusion

This paper proposes a method that consists of two steps: a self-supervised pre-training using denoising diffusion models training objective and a fine-tuning of the obtained model on a radiograph semantic segmentation task. Experiments on dental bitewing radiographs showed that PTDR outperforms baseline self-supervised pre-training methods in the few label regime. Our simple, yet powerful, method allows the fine-tuning phase to easily exploit all the representations learned in the network during the diffusion pre-training phase without any architectural changes. These results highlight the effectiveness of diffusion models in learning representations. In future works, we will investigate the application of this method to other types of medical datasets.

References

1. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv:2112.00390 (2021)
2. Bao, H., Dong, L., Wei, F.: Beit: BERT pre-training of image transformers. arXiv:2106.08254 (2021)
3. Baranchuk, D., Voynov, A., Rubachev, I., Khrukov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022)

4. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Decoder denoising pretraining for semantic segmentation. *Transactions on Machine Learning Research* (2022)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PMLR (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv:2003.04297* (2020)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*. pp. 8780–8794. Curran Associates, Inc. (2021)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems* (2014)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9729–9738 (2020)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* pp. 6840–6851 (2020)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021)
13. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (2021)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. pp. 234–241. Springer (2015)
15. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR* (2015)
16. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning. ICML '08* (2008). <https://doi.org/10.1145/1390156.1390294>
17. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* (2010)
18. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: *International Conference on Medical Imaging with Deep Learning*. pp. 1336–1348. PMLR (2022)
19. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv:2211.00611* (2022)
20. Wu, J., Fu, R., Fang, H., Zhang, Y., Xu, Y.: Medsegdiff-v2: Diffusion based medical image segmentation with transformer. *arXiv:2301.11798* (2023)
21. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: *Proceedings of the European conference on computer vision (ECCV)* (2018)

22. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
23. Xu, Z., Dai, Y., Liu, F., Chen, W., Liu, Y., Shi, L., Liu, S., Zhou, Y.: Swin mae: Masked autoencoders for small datasets. arXiv:2212.13805 (2022)
24. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10145–10155 (2021)

Appendix A: Generalization Capacity

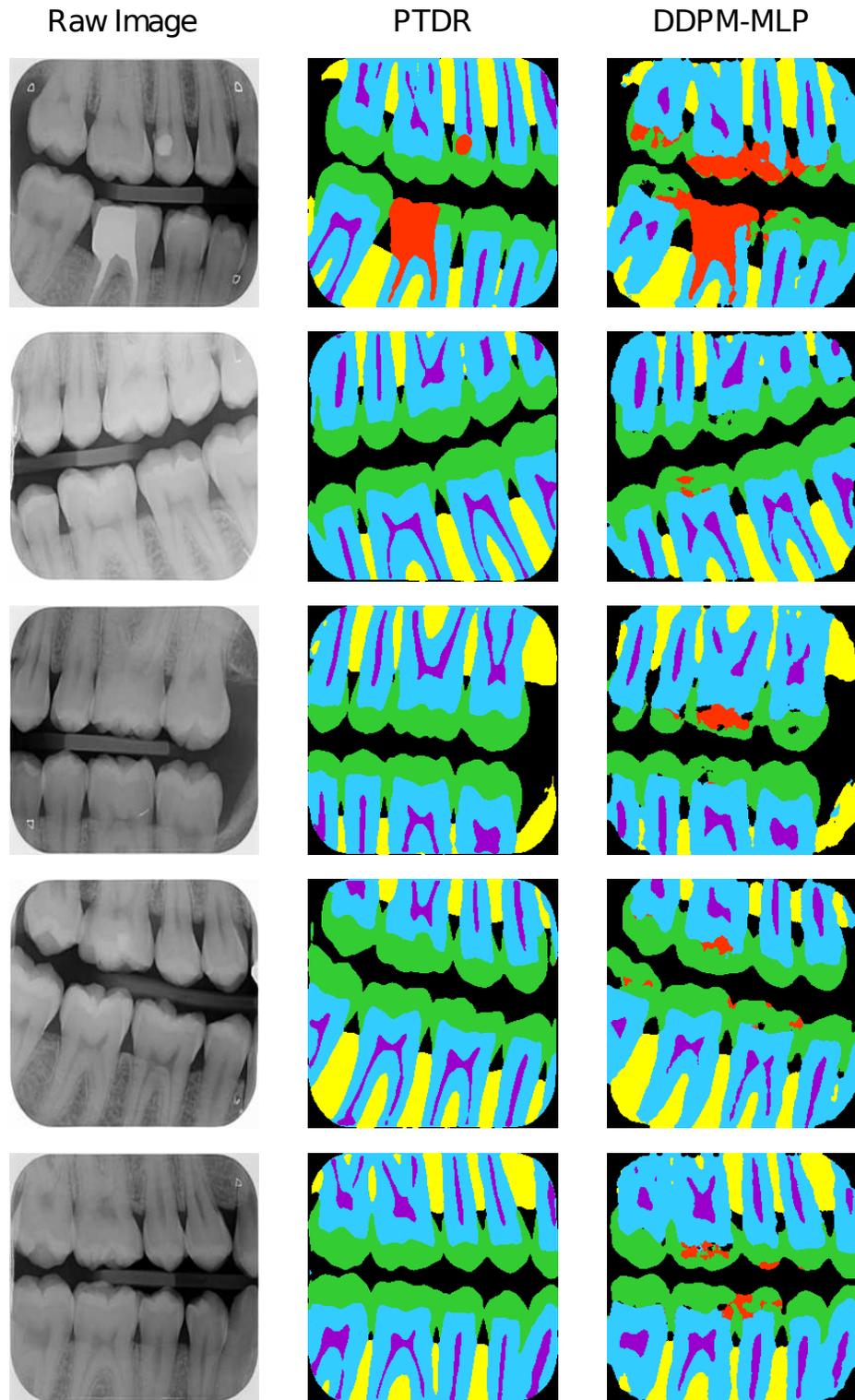
We further tested PTDR on another multi-class semantic segmentation task of lung axial CT images to explore the capacity of our method to transfer to other modalities. For this experiment, Radiopaedia Covid-19 dataset [24] (829 slices) is used for pre-training and COVID-19 CT Segmentation dataset [24] (100 slices) is used for fine-tuning. The latter is annotated for 4 classes: *ground-glass*, *consolidation*, *lung-other* and *background* as semantic maps; and is randomly split with 10 training, 5 validation and 85 test samples. The CT-slices are resized to 256x256, clipped between -1100 and +300 and normalized according to the mean and standard deviation of the clipped Radiopaedia dataset. There is no intersection between the pre-training and fine-tuning dataset. We show that the good performances of the proposed method are not restricted to bitewing radiographs and might be used for other types of medical image segmentation, as illustrated by results of table 3.

Table 3. Performance of PTDR and DDPM-MLP on lung CT images segmentation

Model	Pre-training	mIoU
Unet*	DDPM-MLP [3]	74.64
	PTDR (ours)	81.10

Appendix B: Dataset Generation

We study the ability of the proposed method to generate an artificial dataset. First, a set of dental radiograph is generated using DDPM then a semantic map is generated for each image using both PTDR and DDPM-MLP [3]. We demonstrate qualitatively that the examples generated by PTDR are more consistent than those generated by DDPM-MLP. This capacity paves the way for application such as transfer learning or digital clones.



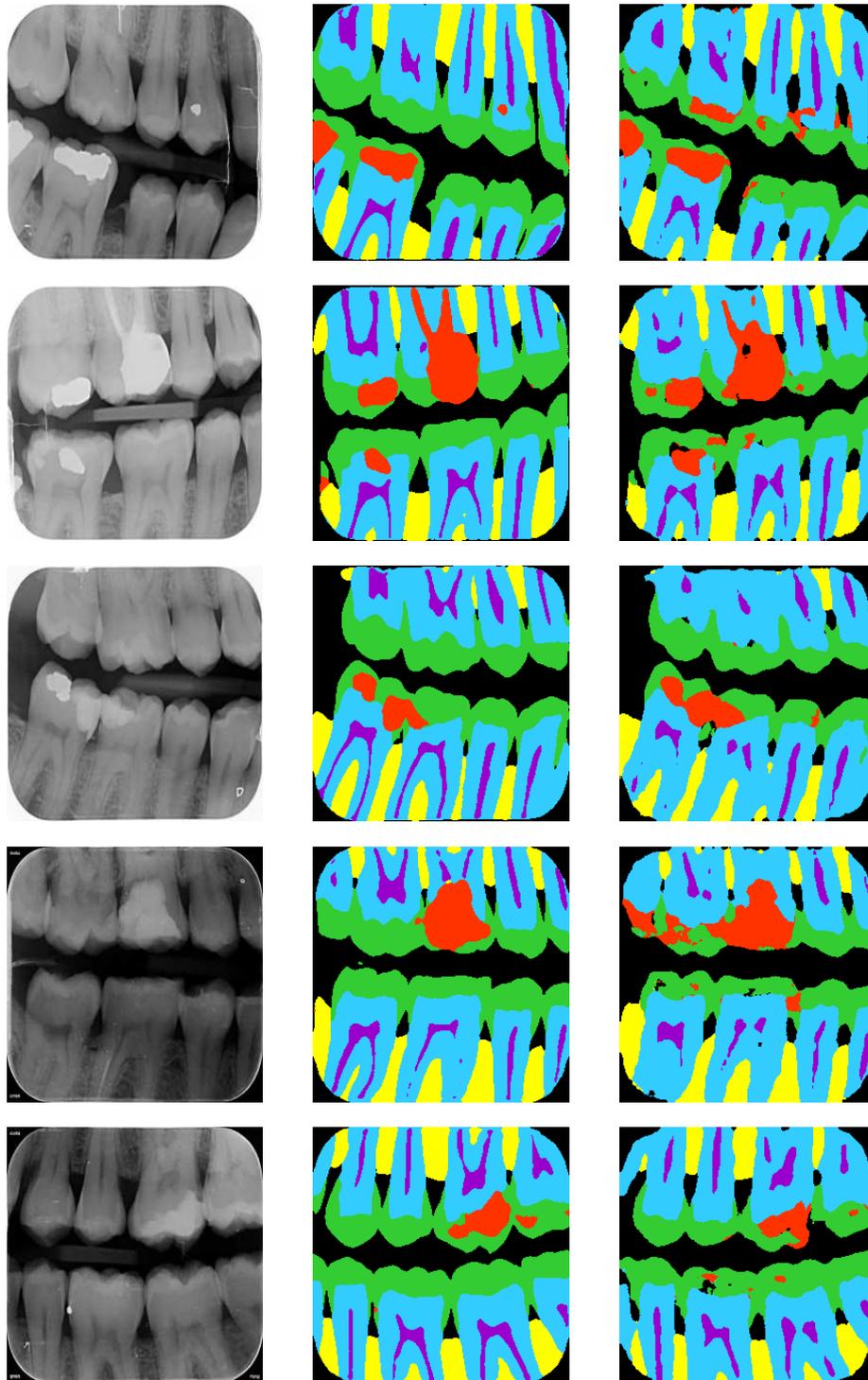


Fig. 6. Generated samples from PTDR and DDPM-MLP