

# Supplementary Materials: PathUp: Patch-wise Timestep Tracking for Multi-class Large Pathology Image Synthesising Diffusion Model

Anonymous Authors

## APPENDIX

### 1 SYNTHESIS OF HIGH-RESOLUTION PATHOLOGY IMAGE

The process of generating high-resolution pathology images, outlined in Alg.1, begins with resizing a genuine pathology image  $I'^{ref}$  to  $512 \times 512$  for use as a spatial context reference input. A controlled level of noise,  $\delta_{t_i}$ , is introduced to create a latent noised reference image, gradually reduced to produce a  $512 \times 512$  pathology image  $I^{ref}$  with similar spatial context. Each  $I^{ref}$  undergoes patch-wise timestep tracking to generate a  $2048 \times 2048$  image, which is then passed through the diffusion model to obtain a high-resolution latent representation. This representation is partitioned into  $N$  patches, merged using weight  $w$ , resulting in a synthetic high-resolution pathology tissue  $I^h$  free of tiling artifacts, yet retaining cancer-related spatial context.

---

#### Algorithm 1 Multi-class large pathology image synthesis

---

```

1: Input: Reference image  $I'^{ref}$ , textural guidance  $c_s$ 
2: Parameter: Latent patch size  $p$ , overlap pixels  $o$ , patch latent weight  $w$ 
3: Output: High-resolution image  $I^h$ 
4:  $x_0'^{ref} \leftarrow \mathcal{E}(I'^{ref})$ 
5:  $x_t'^{ref} \leftarrow \sqrt{\alpha_t} x_0'^{ref} + (1 - \alpha_t) w$ 
6:  $I^{ref} \leftarrow \hat{x}(x_t'^{ref}, c_s)$ 
7: Resize  $I^{ref}$  into  $2048 \times 2048$ 
8:  $X_0 \leftarrow \mathcal{E}(I^{ref})$ 
9:  $X_t \leftarrow \sqrt{\alpha_t} X_0 + (1 - \alpha_t) w$ 
10: Split  $X_t$  into  $N$  patches according to  $p, o$ 
11: for Timestep  $t$  in  $[T, T - 1, \dots, 0]$  do
12:   for Latent patch  $x_t^n$  in  $[1, 2, \dots, N]$  do
13:      $x_{t-1}^n \leftarrow \hat{d}(x_t^n, c_s)$ 
14:   end for
15:   Combine  $x^n$  according to  $w$  for  $X_{t-1}$ 
16: end for
17:  $I^h \leftarrow \mathcal{D}(X_0)$ 
18: return High-resolution pathology image  $I^h$ 

```

---

### 2 DOWNSTREAM TASK

Our evaluation of synthetic high-resolution data entails testing its efficacy in a downstream lesion classification task utilizing the BRACS dataset. Employing the synthetic multi-class high-resolution pathology generated by our model as a data augmentation technique, we integrate it with the training images. Two image classification networks, namely ViT-L and ADMIL, are utilized to assess the performance of our model across both single-instance and

multi-instance learning methodologies. Leveraging the 4465 Regions of Interest (RoIs) provided by BRACS, we segment them into  $2048 \times 2048$  images and conduct a 5-fold cross-validation procedure. Employing a 1:1 generation ratio for all high-resolution images, the models are trained on an Nvidia A100 GPU. For training ViT-L, all multi-resolution images are resized to  $512 \times 512$  with a learning rate of 0.001, while for ADMIL, a learning rate of 0.0001 is utilized. All experiments are trained for 100 epochs.

### 3 EVALUATION METRICS

#### 3.1 Improved Precision and Improved Recall

To calculate Improved Precision (IP) and Improved Recall (IR), given feature vectors representing real images  $\Phi_r$  and generated images  $\Phi_g$ , the nearest neighbor search is conducted for each feature vector  $\phi_g$  in  $\Phi_g$  to find its nearest neighbor  $\phi_r$  in  $\Phi_r$ , and vice versa using the defined binary function  $f_p(\phi, \Phi)$ . Then, IP is computed as:

$$precision(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f_p(\phi_g, \Phi_r) \quad (1)$$

where  $|\Phi_g|$  denotes the cardinality of  $\Phi_g$ , and IR is computed as:

$$recall(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f_p(\phi_r, \Phi_g) \quad (2)$$

where  $|\Phi_r|$  denotes the cardinality of  $\Phi_r$ . These calculations provide insights into how well the generated and real images are aligned in the feature space, indicating the quality of representation between the two image sets.

#### 3.2 Frechet Inception Distance and Kernel Inception Distance

To compute the Frechet Inception Distance (FID) and Kernel Inception Distance (KID), first, synthetic images are generated using the generative model, and feature representations are extracted using a pre-trained Inception-v3 network. For FID, the mean  $\mu_{syn}$  and covariance matrix  $\Sigma_{syn}$  of synthetic image features are compared to those of real images  $\mu_{real}$  and  $\Sigma_{real}$ , using the formula:

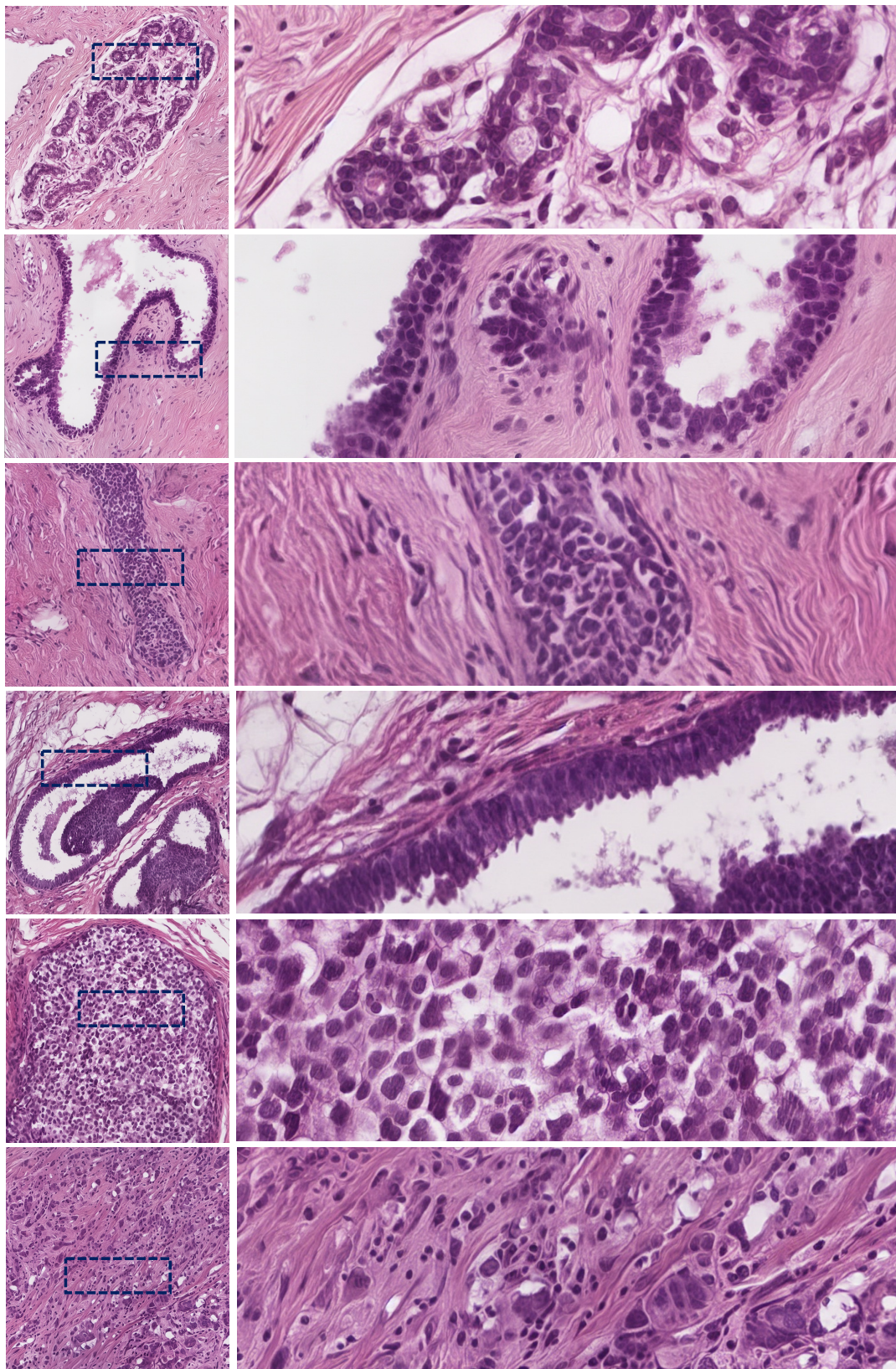
$$FID = |\mu_{syn} - \mu_{real}|^2 + Tr(\Sigma_{syn} + \Sigma_{real} - 2(\Sigma_{syn}\Sigma_{real})^{\frac{1}{2}}) \quad (3)$$

For KID, kernel embeddings are computed for both synthetic and real image features, and the MMD (Maximum Mean Discrepancy) between their distributions is calculated to obtain the KID score, expressed as:

$$KID = MMD^2 = \left| \frac{1}{n_{syn}} \sum_{i=1}^{n_{syn}} \phi(x_{syn,i}) - \frac{1}{n_{real}} \sum_{j=1}^{n_{real}} \phi(x_{real,j}) \right|^2 \quad (4)$$

These metrics offer quantitative assessments of the similarity between distributions of synthetic and real images, aiding in evaluating the quality and fidelity of generated images.





**Figure 1: Examples of multi-class high resolution image generation generated by PathUp. Our method generates realistic high-resolution images with clear tissue such as cell nuclei, glandular tissue and connective tissue. The subtypes are (up to down): Normal, Pathological Benign, Usual Ductal Hyperplasia, Atypical Ductal Hyperplasia, Ductal Carcinoma in Situ and Invasive Carcinoma.**