Conditional Visual Autoregressive Modeling for Pathological Image Restoration

Supplementary Material

1. Metrics

To evaluate the effectiveness of our method in image restoration, we employ four metrics: Peak Signal-to-Noise Ratio (PSNR) [1] and Structural Similarity Index (SSIM) [2] as distortion metrics; Learned Perceptual Image Patch Similarity (LPIPS) [3] and Fréchet Inception Distance (FID) [4] as perceptual metrics. PSNR quantifies the fidelity between a corrected image and a reference image, considering the dynamic range of pixel values. SSIM assesses structural similarity by estimating perceptual differences between images. LPIPS measures perceptual similarity using deep feature representations, where lower values indicate higher visual resemblance. FID evaluates the quality and diversity of generated images by comparing their feature distributions to real images. , with lower scores indicating better alignment with real data.

2. Whole Slide Image Reconstruction

Our proposed Whole Slide Image (WSI) restoration framework employs a latent-space stitching approach that eliminates explicit pixel-level blending. Given an WSI, we first partition it into 32 pixels overlapping 256×256 patches. Because the 256×256 patches can be projected into 16×16 vectors, the shared overlap regions vectors are averaged to compute the optimal latent vectors. Therefore, each patch's prediction can condition on neighboring latent codes and boundary regions can use the adjacent information. Finally, the high-quality indexes of codebook can be fed into the decoder for reconstruction. To show the function of our autoregressive model, we show the reconstruction of WSIs using 8 × 8 vocalbulary index for one 256×256 patch in Figure 1.

3. Downstream Image Classification Tasks

To prove restored images are better than their degraded counterpart in a clinical workflow, we use the tumor detection method DSMIL[5] to recognize the tumor regions in WSI from CAMELYON16. In Figure 2, the results of degraded images exhibited false positives in tumor detection, whereas the restored images provided clearer results and eliminated the false positive issue. Detailed AUC results are shown in Figure 3.

4. Sensitive Analysis

We conduct experiments to find the best vocalbulary size of VQVAE on CAMELYON16 dataset. In Figure 4, we found 4096 is enough for pathological image reconstruction with minor loss compared to larger codebook. When the vocabulary size exceeds 4096, the reconstruction loss (recloss) does not decrease significantly, while the dictionary usage (i.e., the proportion of codewords actually utilized in the codebook) shows a noticeable decline. This indicates that a larger codebook size (e.g., 8192 or higher) does not substantially improve reconstruction quality, as most of the additional codewords remain underutilized.

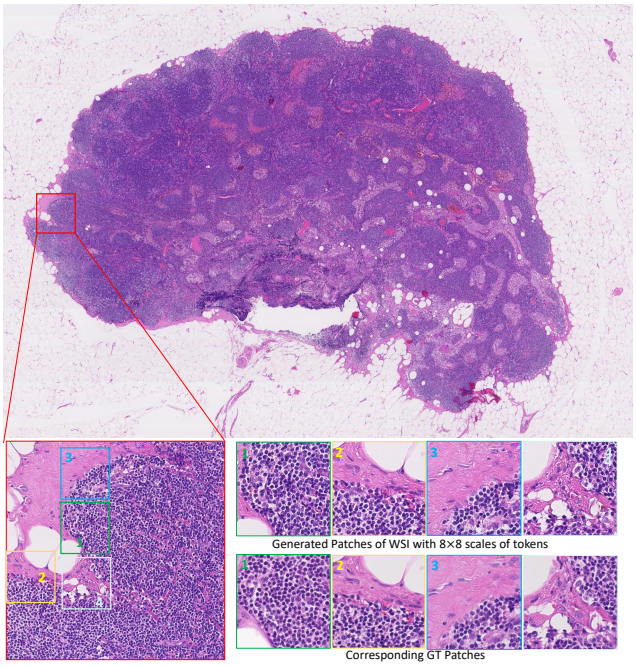

Figure 1. Visual samples of Whole Slide Images (WSIs) using 8 × 8 vocalbulary index for one patch. We predict all the vocalbulary index for 256 × 256 patches and leverage the pretrained decoder to reconstruct the WSIs. Results show the generated images preserve the structure details.

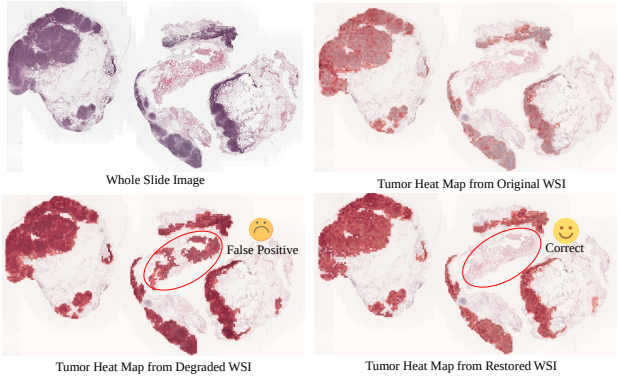

Figure 2. Detection results of DSMIL on original, degraded, and restored WSIs (degradation type: 8× super-resolution).

Methods	AUC
Original	0.8896
Degraded	0.7974
Restored	**0.8922**

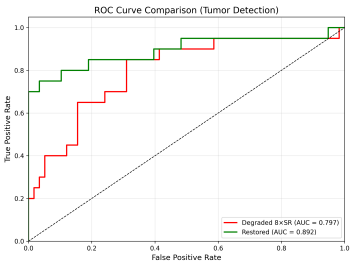

Figure 3. Tumor detection performance comparison: (Left) AUC scores for original, degraded and restored WSIs; (Right) ROC curves for degraded and restored WSIs.

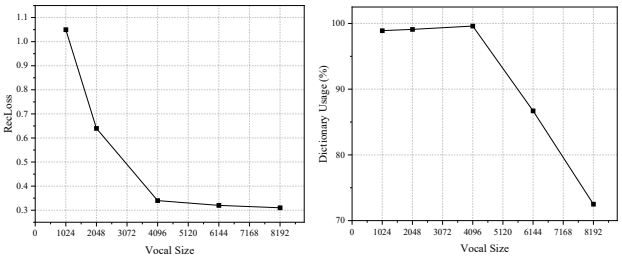

Figure 4. Results of restoration loss (left) and dictionary memory usage (right) under different vocabulary sizes

5. Computational and memory costs

For pathological processing at different resolutions, we measured the GPU memory consumption and inference time per image on an NVIDIA 3090 GPU.

Size	GPU Memory(MB)	Time (sec/image)
256	6,496	0.052
384	7,074	0.109
512	7,838	0.170
768	10,238	0.411
1024	16,662	0.716

6. Coupled Degradation

Real-world pathological images often suffer from coupled degradations (e.g., blur, noise, and downsampling). As shown in Figure 5, CVARPath synthesizes diverse degradation combinations while preserving diagnostically critical features. For instance, a sample labeled (4, 0.3, 7) represents an image that has been downsampled by a factor of 4, corrupted with noise at a density of 0.3, and blurred using a Gaussian kernel of size 7×7.

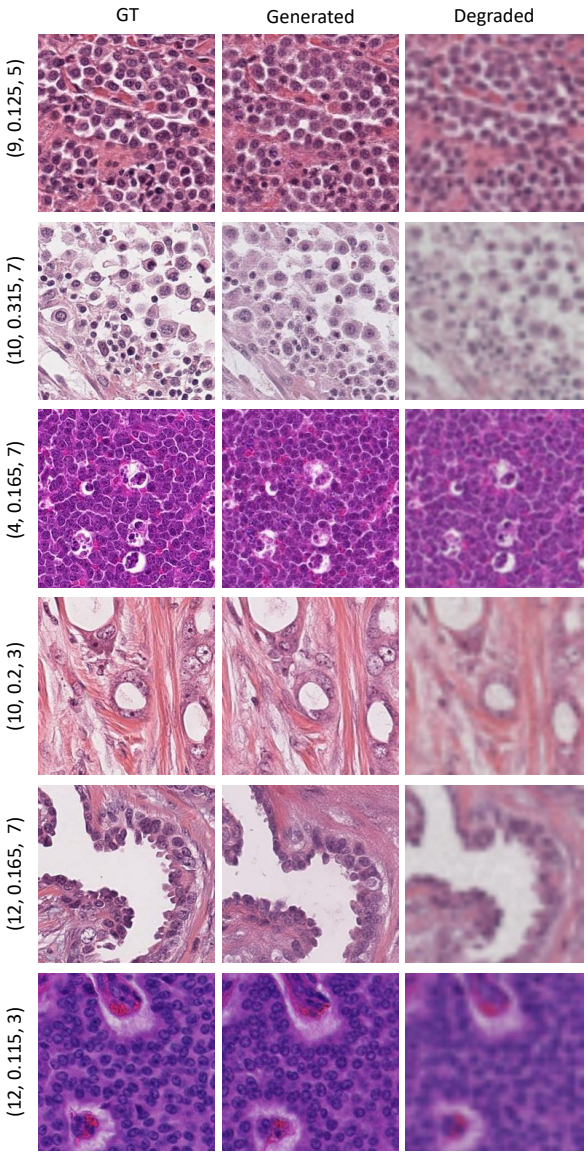

Figure 5. Visual samples of CVARPath for Composite Degradation Tasks. Zoom in for a better view.

References

[1] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 1

[2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[5] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 1