ZoomLDM: Latent Diffusion Model for multi-scale conditional histopathology image generation

Supplementary material

5. Effectiveness of conditioning mechanism

Our conditioning mechanism consists of a CNN trained jointly with the diffusion model, designed to handle large embedding matrices efficiently. The CNN is structured to transform $128 \times 128 \times 384$ SSL embeddings into an output tensor of $8 \times 8 \times 512$.

We probe the compression capabilities of the summarization CNN by comparing the performance of multiple-instance learning (MIL) algorithms [5, 6] using features extracted from our CNN and HIPT [7] at $20\times$ magnification. On average, each WSI contributes two patches at the $0.15625\times$ magnification level. Processing these patches through the CNN yields a $8 \times 8 \times 512$ tensor for each, culminating in a 64×512 representation for each WSI. In comparison, applying the HIPT method to $20\times$ magnification patches results in each WSI represented by a set of feature vectors totalling in size 8000×384.

In Table 2, we present the results of training the MIL algorithms on the full dataset (100 %) and a reduced subset (25 %). To ensure consistency, use a 10-fold cross-validation strategy, aligning with the data splits from HIPT. The results indicate that the CNN features closely match and, in one scenario, even surpass the performance of HIPT features. This observation is noteworthy for two reasons: (i) the CNN features contain $45\times$ less information than the HIPT features and (ii) the CNN training did not involve a contrastive loss objective. This highlights that through learning to synthesize images, the model inherently acquired the ability to discriminate, becoming skillful at extracting the essential information from SSL embeddings, eliminating redundancies and merging self-supervised knowledge with generative capabilities.

Feature source	# features per WSI	Feature Dimension	x times info compression	25% training		100% training	
				CLAM-SB	DSMIL	CLAM-SB	DSMIL
HIPT	8000	384	1x	0.788	0.784	0.861	0.839
Emb CNN	128	512	45x	0.754	0.709	0.878	0.805

Table 2. Performance comparison of MIL algorithms using $0.15625\times$ CNN features and $20\times$ HIPT features. The CNN features not only match but occasionally outperform HIPT features despite being $45\times$ more compact, underscoring the CNN's learned ability to extract essential information.

6. Discussion

Whole slide image (WSI) classification analysis often involves extracting patch-level features using a pretrained SSL encoder [1, 3, 8], followed by a Multiple Instance Learning (MIL) framework [4, 5]. This method's weakly supervised nature typically necessitates the use of extensive datasets, often requiring hundreds of thousands of WSIs for effective training [2]. The scale of data needed poses significant challenges in terms of storage, processing, and analysis, highlighting a critical bottleneck in current methodologies for handling histopathological data at scale.

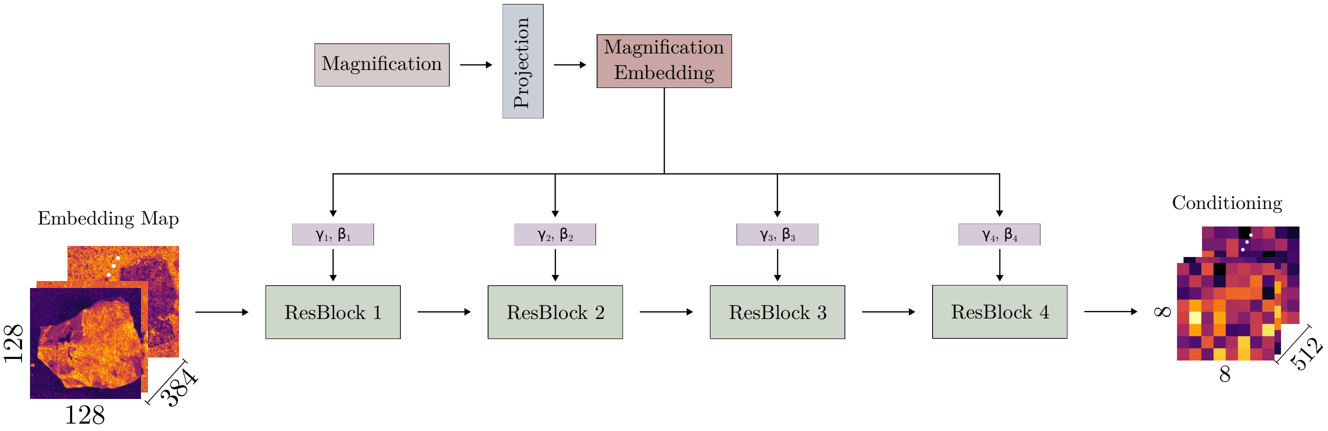

Figure 3. Architecture of the summarization CNN. The CNN is jointly trained with the diffusion model, and is designed to efficiently process and condense SSL embeddings.

Our CNN condenses information into a 65536-dimensional vector for each WSI. These vectors, roughly 50 kb each, can streamline the storage and analysis of extensive WSI datasets, greatly simplifying data management challenges. Local processing of WSIs into compact vectors can simplify the exchange of crucial data between institutions without the usual logistical and privacy concerns. Such a practical solution fosters improved collaboration and makes histopathological research and diagnostics more efficient, demonstrating the practical benefits of training diffusion models effectively across various magnifications.

We also investigate how we could perform WSI compression with the proposed Latent Diffusion Model and the SSL conditioning. Instead of storing whole-slide images at $20\times$ magnification, which usually are $\approx 50k \times 50k$ pixels in resolution, we can store the extracted representations and regenerate the necessary parts of the image at will, at all magnifications, resulting in roughly $50\times$ compression of WSIs. In Figure 4 we demonstrate a decompression pipeline, where we re-synthesize patches at all magnifications of the image, stored as SSL embeddings. The decompression is performed by initializing the diffusion process at a single magnification from an intermediate timestep ($t = 500$), using an upsampled version of the previous magnification patch.

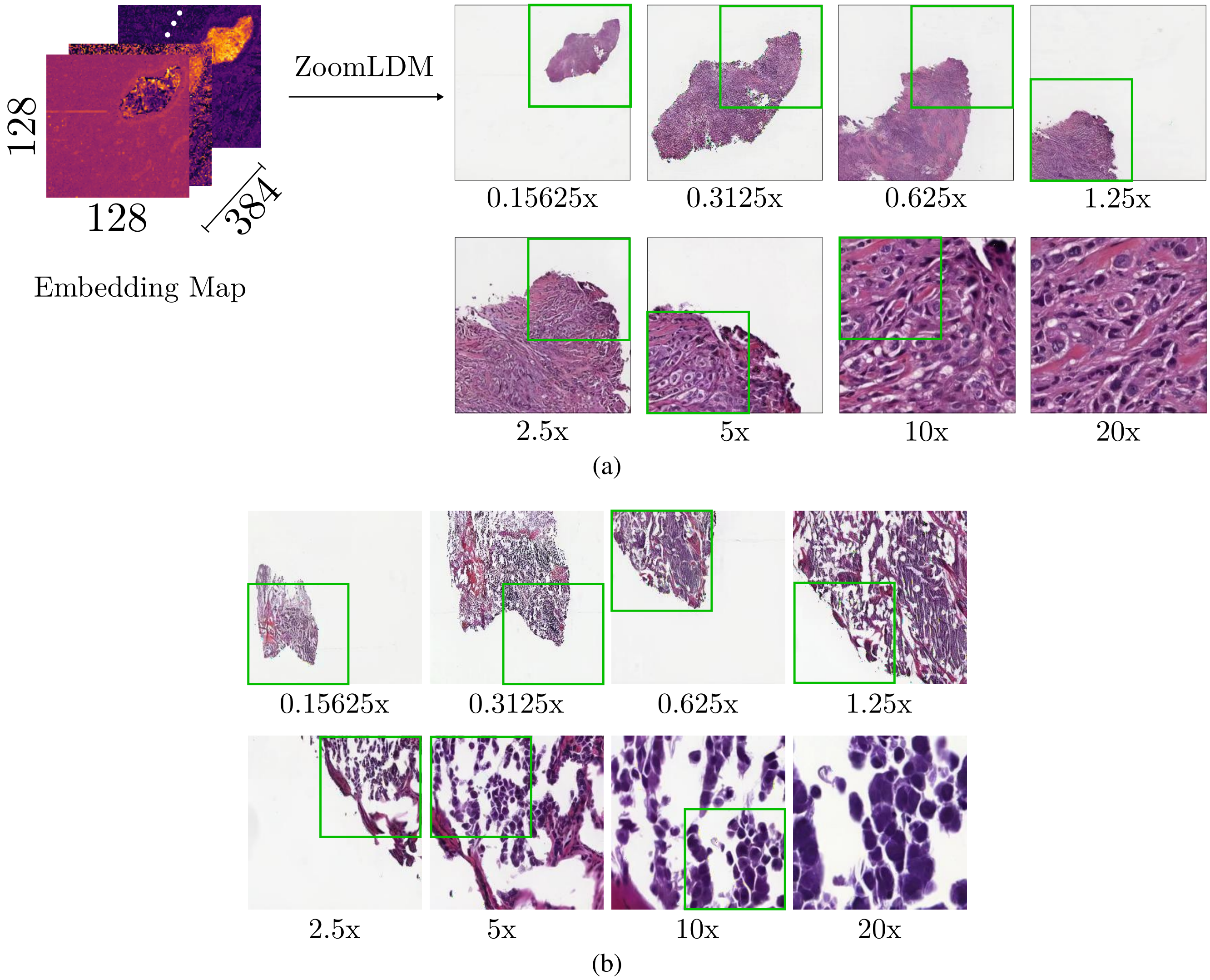

Figure 4. Image “decompression” from a small embedding map representation to all magnifications using ZoomLDM.

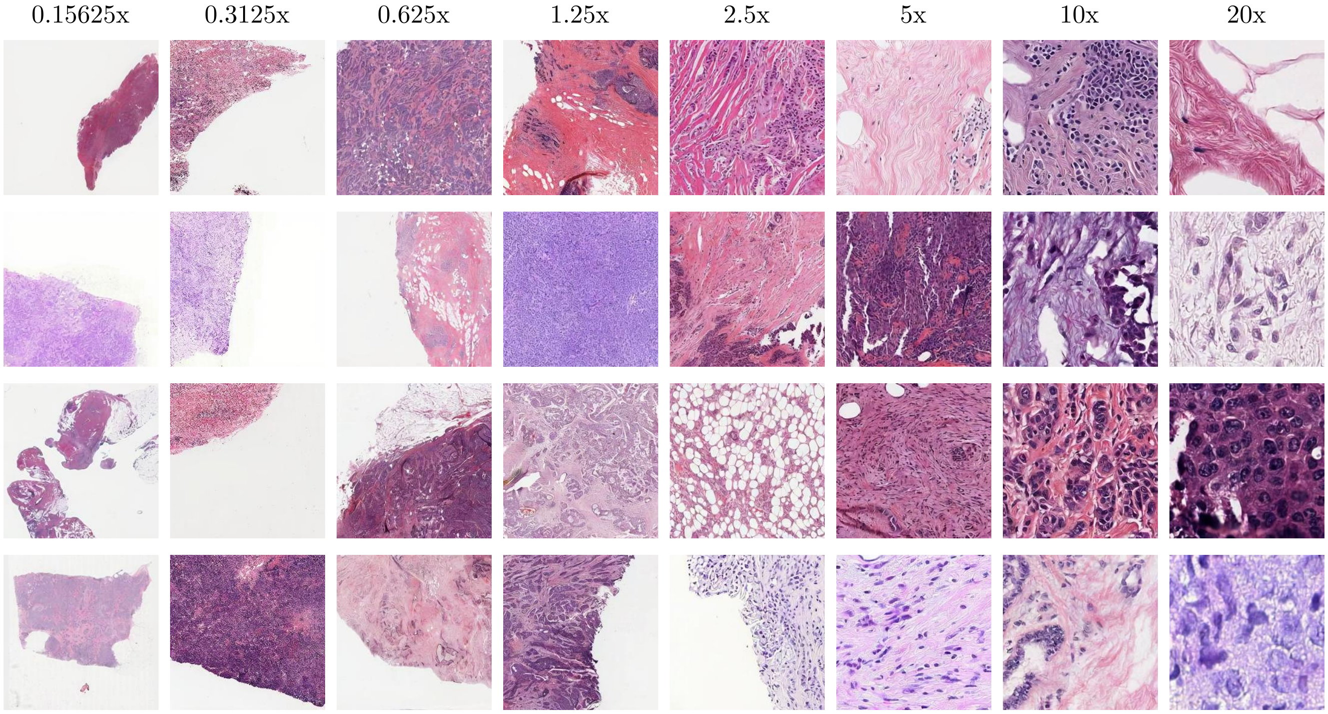

Figure 5. ZoomLDM-syntheszed images using conditions sampled from the embedding diffusion model.

References

[1] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1

[2] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023. 1

[3] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 1

[4] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 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

[6] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570, 2021. 1

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[8] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. 1