DISC: Latent Diffusion Models with Self-Distillation from Separated Conditions for Prostate Cancer Grading

Man M. Ho[1] Elham Ghelichkhan [1] Yosep Chong[2,4] Yufei Zhou[3]
Beatrice Knudsen[4,5] Tolga Tasdizen[1,6]
[1] Scientific Computing and Imaging Institute, University of Utah, Utah, USA
[2] The Catholic University of Korea College of Medicine, Seoul, Korea
[3] Case Western Reserve University, Ohio, USA
[4] Departmant of Pathology, University of Utah, Utah, USA
[5] Huntsman Cancer Institute, University of Utah Health, Utah, USA
[6] Department of Electrical and Computer Engineering, University of Utah, Utah, USA

Abstract

Latent Diffusion Models (LDMs) can generate high-fidelity images from noise, offering a promising approach for augmenting histopathology images for training cancer grading models. While previous works successfully generated high-fidelity histopathology images using LDMs, the generation of image tiles to improve prostate cancer grading has not yet been explored. Additionally, LDMs face challenges in accurately generating admixtures of multiple cancer grades in a tile when conditioned by a tile mask. In this study, we train specific LDMs to generate synthetic tiles that contain multiple Gleason Grades (GGs) by leveraging pixel-wise annotations in input tiles. We introduce a novel framework named Self-Distillation from Separated Conditions (DISC) that generates GG patterns guided by GG masks. Finally, we deploy a training framework for pixel-level and slide-level prostate cancer grading, where synthetic tiles are effectively utilized to improve the cancer grading performance of existing models. As a result, this work surpasses previous works in two domains: 1) our LDMs enhanced with DISC produce more accurate tiles in terms of GG patterns, and 2) our training scheme, incorporating synthetic data, significantly improves the generalization of the baseline model for prostate cancer grading, particularly in challenging cases of rare GG5, demonstrating the potential of generative models to enhance cancer grading when data is limited.

1. Introduction

In recent years, Latent Diffusion Models (LDMs) [28, 30] have emerged as a powerful tool in computational pathology for generating high-fidelity tiles for Whole Slide Images (WSIs). Synthetic histopathology images potentially improve multiple downstream tasks, one of which is the training of cancer grading models (please refer to Supplementary Document for application overview). However, in prostate cancer grading, the current utilization of LDMs remains limited, as it does not effectively incorporate multiple cancer grades into synthetic tiles. In prostate cancer, the growth pattern of the cancer is used to define the cancer grade, which is named the Gleason grade. Pathologists identify 5 different Gleason Grade (GG) groups to forecast the severity of prostate cancer and likelihood of cancer progression. However, synthetic tiles generated by LDMs may display incorrect Gleason grade patterns, or mistake benign

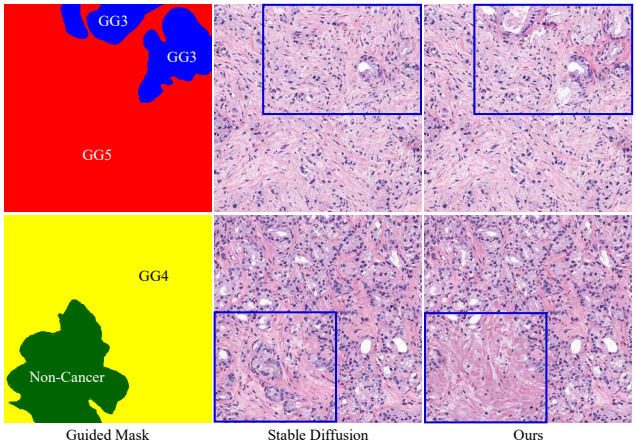

Figure 1. Stable Diffusion [30] produce a sheet of cells resembling GG5 in GG3-indicated regions (*top*) and fused glands resembling GG4 in Non-Cancer-indicated regions (*bottom*).

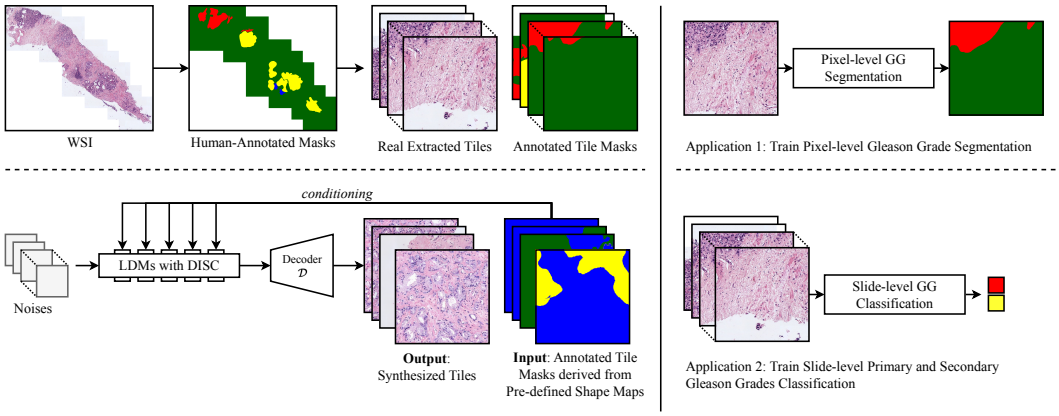

Figure 2. Besides the real patches (*top-left*) for training pixel-level and slide-level Gleason grading models (*right*), we introduce Latent Diffusion Models (LDMs) [30] with Self-**Di**stillation from **S**eparated **C**onditions (DISC) to accurately generate admixtures of multiple Gleason Grades in a tile when conditioned by a tile mask (*bottom-left*).

glands for high grade cancer (Fig. 1). To address these issues, we introduce a novel approach: we first tailor LDMs to produce tiles conditioned by human-annotated masks that feature multiple GG labels. Building on the principle of "*Get More Done - One Thing at A Time*" [43], we further refine this approach with our Self-Distillation from Separated Conditions (DISC) technique, aimed at improving the precision of GG patterns guided by intricate masks. Leveraging the methodology outlined in [16], we also develop a training framework that efficiently utilizes generated tiles to enhance the performance of both pixel-level and slide-level cancer grading models, as illustrated in Figure 2. Moreover, we implement a straightforward yet effective sampling strategy to ensure a balanced representation of GGs within the tile masks, thus addressing potential label distribution imbalances in the training dataset. Our work is available at https://minhmanho.github.io/disc/.

Advancements in Histopathology Image Synthesis. Following the success of Generative Adversarial Networks (GANs) [4, 12, 21, 22] in image synthesis, Diffusion Models have become a leading approach for generating high-fidelity images from noise [15, 32, 36, 37, 45]. Rombach et al. [30] have significantly advanced the field with the introduction of Latent Diffusion Models (LDMs), which demonstrate exceptional image synthesis capabilities with reduced computational demand by utilizing pre-trained autoencoder-based latent spaces. These innovative generative models [28] are revolutionizing computational pathology by providing robust data augmentation capabilities for a variety of downstream applications, including nuclei segmentation [6, 10, 17], polyp segmentation [38], the analysis of skin lesions, and the classification of Renal Cell Carcinoma (RCC) [8]. In our work, we specifically tailor LDMs to generate image tiles guided by complex masks that incorporate multiple Gleason Grades (GGs). Furthermore, we introduce Self-Distillation from Separated Conditions (DISC), an innovative method aimed at improving the precision of label patterns in the guided mask. Through the training of pixel-level and slide-level cancer grading models, such as Carcino-Net [25] and TransMIL [34], alongside our synthetic tiles, we observe significant performance improvements, especially in diagnosing rare cases like GG5.

Knowledge Distillation (KD) for Generative Models. KD is a technique that transfers knowledge from a larger, more complex model (teacher) to a smaller, simpler model (student) [14]. In image classification, self-distillation, introduced in [44], distills knowledge from deeper classifiers to shallower ones in neural networks. Self-distillation with no labels (DINO) [7] employs co-distillation [1] to enhance the performance of Vision Transformers [11]. KD is also used in GAN-based image synthesis to improve results and computational efficiency [24, 41, 42]. For example, Self-distilled StyleGAN [29] filters uncurated internet images using a pre-trained StyleGAN and fine-tunes the model to generate images closer to cluster centers defined by the latent space. Meanwhile, KD has been applied for LDMs to improve sampling efficiency [27, 33]. Inspired by [43], we then separate the mask into single label masks and denoise latent features with one mask at a time, resulting in higher-confidence patterns for labels indicated in the label-guided mask. Finally, we propose Self-Distillation from Separated Conditions (DISC) to optimize computational cost while improving the quality of generated patterns.

Our contributions are as follows: 1) We propose the application of Latent Diffusion Models (LDMs) to generate histopathology patches using guided masks with multiple Gleason Grades. 2) We address the issue of LDMs generating incorrect labels when complex masks are provided by introducing Self-Distillation from Separated Conditions (DISC). 3) Our work surpasses previous studies in

two aspects: (a) LDMs with DISC produce more accurate histopathology images compared to LDMs [30]. (b) Training baseline models such as Carcino-Net [25] and TransMIL [34] with our generated tiles leads to significant improvements on both in-distribution SICAPv2 [35] and out-of-distribution LAPC [23] and PANDA [5] datasets, particularly for the rare case of Gleason Grade 5 with limited data. This highlights the potential of generative models in enhancing rare cancer grading/detection with limited data.

2. LDMs with DISC for Cancer Grading

Latent Diffusion Models (LDMs) have shown their capability of generating high-fidelity images from noises, creating a promising approach for augmenting histopathology images in training cancer grading models. Although the previous works can generate high-fidelity histopathology images using LDMs, generating histopathology images with multiple Gleason Grades (GGs) is not entirely exploited, and the utilization of these generated images to improve the downstream task like pixel-level and slide-level Prostate Cancer (PCa) Grading is still an open question. Besides, LDMs still suffer from generating histopathology images conditioned by complex masks, as shown in Figure 1. In this work, we present specific LDMs, which can generate multiple GGs by leveraging pixel-wise annotation masks, discussed in Section 2.1. For slide-level cancer grading models that require training pairs as {multiple tiles, primary and secondary GGs}, we employ an efficient sampling technique. This strategy enables the generation of tile sets tailored to specific primary and secondary Gleason grades, as detailed in Section 2.2. Additionally, to address the limitations of LDMs, we introduce the Self-Distillation from Separated Conditions (DISC) method, aimed at producing more precise GG patterns in alignment with GG-guided masks, as explored in Section 2.3. Lastly, we design a training framework that employs the generated tiles to significantly improve the accuracy of existing pixel-level and slide-level cancer grading models, as detailed in Section 2.4. The comprehensive process is illustrated in Figure 2.

2.1. LDMs conditioned by Gleason Grades

In this study, we train a generative model based on Latent Diffusion Models (LDMs) [30] (as known as Stable Diffusion) that is conditioned by guided masks with multiple Gleason Grades (GGs). Specifically, we consider an image tile $x \in \mathbb{R}^{H\times W\times 3}$ along with its pixel-wise annotated mask $m \in \{0,1,2,3\}^{H\times W}$, where the labels $0,1,2,3$ represent for Non-Cancer, GG3, GG4, and GG5, respectively. To extract essential features and reduce noises for high-quality image synthesis, we utilize a pre-trained VQ-regularization auto-encoder [39] with encoder $\mathcal{E}$ and decoder $\mathcal{D}$ provided by [30] to encode and downsample the input image x into a latent representation $z = \mathcal{E}(x)$, that $z \in \mathbb{R}^{h\times w\times c}$, with

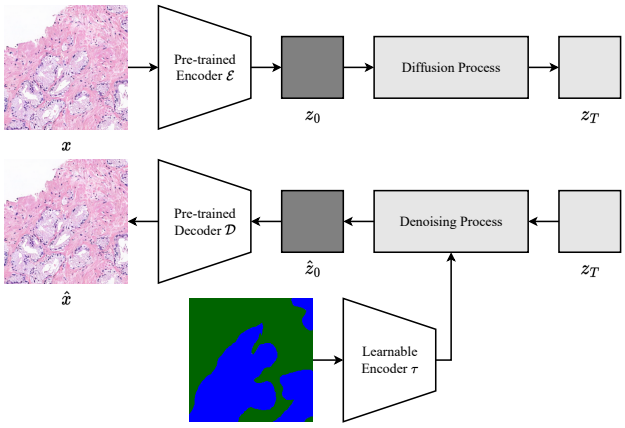

Figure 3. Latent Diffusion Models [30] conditioned by guided masks with multiple Gleason Grades (GGs)

a factor $f = H/h = W/w = 4$. Subsequently, the decoder D reconstructs the latent z back to the input image $\hat{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, as shown in Figure 3. In line with [30], we employ a denoising U-Net [31], denoted as ϵ_θ, to estimate the Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. This denoising model is conditioned by a GG-guided mask m, which is preprocessed by convolutional-layer-based τ_θ. Concretely, m is fed to the encoder and decoder layers via cross-attention layers [18, 19, 30, 40]. The objective is to minimize the simplified loss function:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,m,\epsilon\sim\mathcal{N}(0,I),t}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(m))||_2^2]. \quad (1)$$

Here, ϵ_θ and τ_θ are jointly optimized, and z_t represents the noisy version of z_0 at time step t, sampled uniformly from $\{1, ..., T\}$, where $T = 1000$.

2.2. Tile Annotation Mask Sampling

After training the denoising U-Net to predict added Gaussian noise accurately, we employ the DDIM sampler [30, 37] for faster sampling of image tiles conditioned by multiple Gleason Grades (GGs) with $T_{DDIM} = 200$. To obtain and augment the annotation shapes, we preprocess existing human-annotated masks m from SICAPv2 [35], converting them into tile shape masks, denoted as $m_{freq} \in \{0,1,2,3\}^{H\times W}$. Here, labels are reclassified according to their frequency distribution, from the most to the least frequent (0-to-3). This approach aims to preserve the annotators' drawings, thus generating more authentic-looking tiles at low cost. In sampling phase, a tile shape map is randomly selected and non-overlapping cancer grading labels are assigned based on the random weights, which determines the label majority for a large number of m_{freq}. In simulating a tile set for slide-level classification via Multiple Instance Learning (MIL), we randomly select 20-100 annotation shape masks per designated primary and secondary GGs, with non-overlapping cancer grading labels

distributed according to random weights. When the primary GG is Non-Cancer, the Non-Cancer label is exclusively applied. A comprehensive explanation and ablation study on Random Weights are in Supplemental Document.

2.3. Self-Distillation from Separated Conditions

While Latent Diffusion Models (LDMs) [30] are capable of producing high-fidelity tiles specifically designed for particular primary and secondary Gleason Grades (GGs) post-training, they still encounter difficulties in accurately generating Gleason patterns with high confidence for designated areas within the GG-guided mask. For example, when conditioned on pixel-wise human-annotated masks, LDMs might inaccurately generate a sheet of cells representing GG5 patterns in areas marked for GG3, where glandular structures are expected. Also, the Non-Cancer pattern generated by LDMs exhibits fused glands representing GG4 instead of stroma or uniform glands, as shown in Figure 1.

To address these issues, we draw inspiration from the characteristic of LDMs, which can generate high-confidence patterns for a single GG throughout the entire denoising process. We propose a denoising process with Separated Conditions (SC) for LDMs. At the start of the denoising process, we duplicate a Gaussian noise $z_T \sim \mathcal{N}(0, I)$ K times to obtain a collection $\{z_T^0, \ldots, z_T^{K-1}\}$, where $K = 4$ represents the number of labels. Subsequently, LDMs denoise and infer $\{z_0^0, \ldots, z_0^{K-1}\}$ at time step $t = 0$ from $\{z_T^0, \ldots, z_T^{K-1}\}$ conditioned by the corresponding separated masks $sm_0, \ldots, sm_{K-1}$, where $sm_k \in \{k\}^{H \times W}$ and $k \in \mathbb{Z}, \quad 0 \leq k < K$. This enables the generation of z_0^k with the strong characteristic patterns of label k. To generate the final latent feature representing the guided complex mask m, we downsample m using nearest-neighbor interpolation and separate it into binary masks $m_0, \ldots, m_{K-1}$, where $m_k \in 0, 1^{h \times w}$ denotes the regions corresponding to label k in m. We then multiply the latent features z_0^k with the binary masks m_k and merge them together to generate the final latent representation z_0^{mixed} as:

$$z_t^{mixed} = Fuse(z_t^k, m_k) = \sum_{k=0}^{K} z_t^k \cdot m_k \tag{2}$$

While the denoising process with SC enhances GG patterns in the generated tiles, it also increases time complexity by a factor of K. To maintain the speed of the vanilla denoising process conditioned by a complex mask, we propose Self-Distillation from Separated Conditions (DISC), where vanilla denoising process can mimic latent features z_t^{mixed} from the denoising process with SC by optimizing $||z_t^{mixed} - z_t||_1$, as illustrated in Figure 4. To improve the efficiency of fine-tuning LDMs with DISC, we retain only the final latent features z_0^k and define a simplified loss:

$$\mathcal{L}_{DISC} = \mathbb{E}_{z_0^{mixed}, m, \epsilon \sim \mathcal{N}(0,I), t}[||\epsilon - \epsilon_\theta(z_t^{mixed}, t, \tau_\theta(m))||_2^2] \tag{3}$$

where the noisy z_t^{mixed} can be obtained using a cummulative noise scheduler $\overline{\alpha}$ [15] as $z_t^{mixed} = \sqrt{\overline{\alpha}_t} z_0^{mixed} + \sqrt{1 - \overline{\alpha}_t}\epsilon$. Here, the generation of z_0^{mixed} from z_0^k with any random mask m is achieved through Equation 2. Eventually, we provide four models for further evaluation: 1) **SD**: Latent Diffusion Models (LDMs) [30], also known as Stable Diffusion (SD), for generating tiles from a WSI conditioned by guided masks (top of Figure 4), 2) **SD-SC**: The pre-trained SD generates tiles with Separated Conditions (bottom of Figure 4), 3) **SD-DISC**: We generate $20,000$ samples of separated z_0^k and continue to fine-tune the pre-trained SD exclusively on these samples, optimizing the loss function $\mathcal{L}_{DISC}$ from Equation 3 (top+bottom of Figure 4), and 4) **SD-DISC-CoTrain**: We also train SD-DISC with real data. This involves averaging the training errors from both $\mathcal{L}_{LDM}$ and $\mathcal{L}_{DISC}$.

2.4. Training Prostate Cancer Grading Models

We demonstrate the effectiveness of our scheme in improving pixel-level and slide-level cancer grading performance of existing models by training CarcinoNet [25] and TransMIL [34] on both real and synthesized tiles from the SICAPv2 dataset [35]. In slide-level grading, which involves predicting primary and secondary Gleason Grades (GGs), we modify the last layer of TransMIL from multi-class classification with a Softmax function to multi-label classification with a Sigmoid function. The TransMIL model is trained by minimizing the following loss function: $\mathcal{L}_{slide} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})$. Here, y denotes the ground-truth label, and $\hat{y}$ represents the predicted label. To analyze the impact of generated histopathology images on improving the slide-level cancer grading model, inspired by [16], we set a Balance Weight $\lambda \in [0, 1]$ to balance training errors between real and synthesized samples:

$$\mathcal{L}_{total} = (1 - \lambda)\mathcal{L}_{real\_slide} + \lambda \mathcal{L}_{synthesized\_slide} \tag{4}$$

A higher value of λ indicates a greater emphasis on optimizing the model using synthesized samples.

3. Experiments

In this section, we discuss the training and assessment of Latent Diffusion Models (LDMs) [30] using our proposed Self-Distillation from Separated Conditions (DISC) technique. Subsequently, we perform an ablation study on LDM conditions, including tile-level and pixel-level labels (please refer to the Supplementary Document for layouts [45]). To

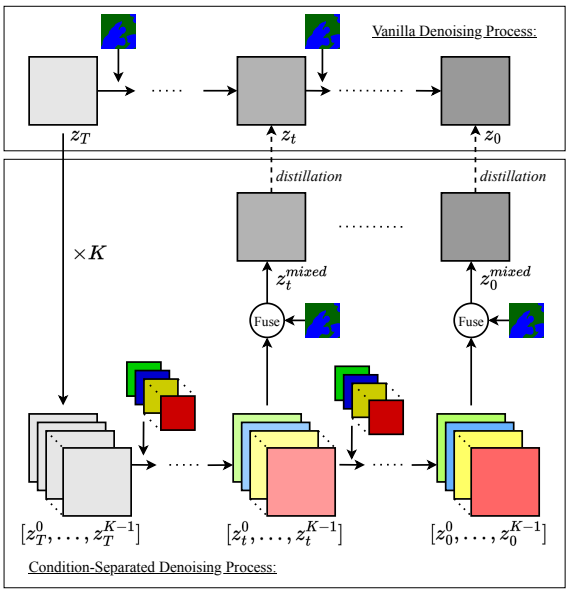

Figure 4. We introduce Self-**Di**stillation from **S**eparated **C**onditions (DISC) to improve image synthesis accuracy. Instead of using the initial complex guided mask with multiple Gleason Grades (GGs) (**top**), we generate separate latent features with distinct labels, which are fused with the mask in the final step for robust patterns. However, this approach incurs a computational cost of $\times K$, the number of labels. To address this, we train the main process to distill information from fused latent features obtained from the Condition-Separated Denoising Process (**bottom**).

establish our approach's superiority, we present a qualitative comparison among various models: vanilla Stable Diffusion (SD) [30], SD with Separated Conditions (SD-SC), SD fine-tuned with DISC using 20,000 generated separated samples (SD-DISC), and SD-DISC fine-tuned with actual tiles (SD-DISC-CoTrain), all outlined in Section 2.3. Notably, we focus on qualitative assessment instead of relying on quantitative evaluation metrics such as FID and Inception Score (IS), which may not capture generated images with incorrect patterns. Moreover, we highlight the effectiveness of our generated data in enhancing both pixel-level and slide-level prostate cancer grading performance. To achieve this, we train and compare the baseline models to those that have been jointly trained with tiles generated by our ablation models. Specifically, we utilize CarcinoNet [25] as our baseline for pixel-level classification. For slide-level classification, TransMIL [34] serves as the baseline, and Mixed Supervision [3] is used as a comparison model. Given that pixel-level annotations can be imprecise and incomplete [2], we qualitatively assess the segmentation results presented in this paper. Additionally, we perform a quantitative evaluation of the precision of pixel-level classification models, as detailed in Supplementary Document. For slide-level cancer grading, the models are evaluated using the Area Under the Receiver Operating Characteristic Curve (AUCROC) for the multi-label classification task of prostate cancer grading. Furthermore, we investigate how synthesized histopathology images affect the models' generalization by adjusting the balance weight λ within the range of $[0.0, 0.9]$. All experiments are conducted on an NVIDIA RTX A6000 GPU.

Training Latent Diffusion Models. Firstly, we train SD [30] on two folds of the SICAPv2 dataset [35], where each fold has approximately 96 WSIs (7500 tiles) and 28 (2500 tiles) for training and validation, respectively. Following the completion of this dual-fold training phase (spanning 7 days), we select the model with the lowest validation error across both folds and transition to training it on the entire training dataset. This extended training phase for the chosen model spans 50 epochs, ensuring optimal generalization. Once SD is proficiently trained to generate prostate tiles, we then prepare 20,000 samples z_0^k from Gaussian noise for fine-tuning using the Self-Distillation from Separated Conditions (DISC) technique. From this point, there are two pathways for further fine-tuning SD: (1) Fine-tuning on the 20,000 generated samples using DISC, denoted as SD-DISC, and (2) Jointly fine-tuning on both real tiles and the 20,000 generated samples, denoted as SD-DISC-CoTrain. It is important to note that SD-SC does not require additional training, as it utilizes the already well-trained SD model to generate condition-separated latent features. These pathways cost approximately 2-3 days.

Training and evaluating cancer grading models. We trained the TransMIL model [34] and our ablation models using pre-extracted image tiles across 4 folds provided by the SICAPv2 dataset [35]. Concurrently, Mixed Supervision [3] employs a method of extracting tiles based on superpixel regions with centroid coordinates, similar to SegGINI [2]. This strategy ensures more reliable instance-level labels, as patterns within the same region are more similar. In our study, beyond utilizing existing tiles, we generated 276 tile sets representing 276 whole-slide images (WSIs) to balance the SICAPv2 dataset. The generation of a tile set, comprising 20-100 tiles, ranges from 4 to 16 minutes. The number of Whole Slide Images (WSIs) for each primary Gleason Grade (GG) was increased by generating additional tile sets as WSIs using models like SD, SD-SC, SD-SC-DISC, and SD-DISC-CoTrain, resulting in a complete set of 100 WSIs for every grade. We maintained consistency in tile generation for fair comparison by setting specific random seeds, which influenced the selection of GG-guided masks and Gaussian noise. For evaluation purposes, we not only utilized the test samples from the SICAPv2 dataset but also prepared a balanced test set with 100 WSIs for each label from out-of-distribution PANDA dataset [5]. Additionally, we assessed the pixel-level performance of CarcinoNet using $2,200$ tiles from the LAPC dataset [23]

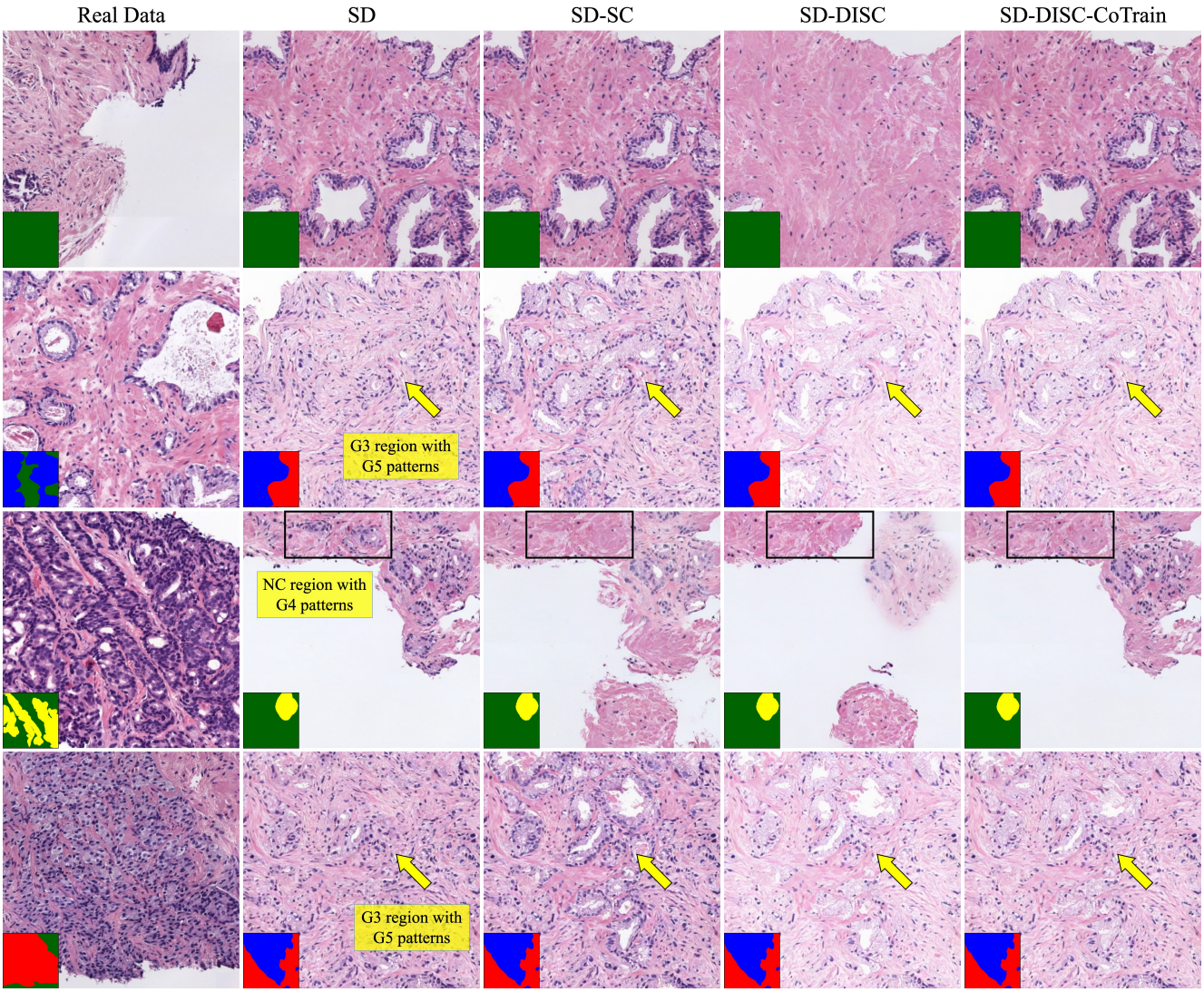

Figure 5. A qualitative comparison between Stable Diffusion (SD) [30] and our proposed technique, SD with Self-Distillation from Separated Conditions (DISC) (discussed in Section 2.3), for histopathology image synthesis. This work yields higher-confidence label patterns compared to SD. Notably, SD tends to generate fused glands representing GG4 for Non-Cancer regions (highlighted rectangles) and sheets of cells representing GG5 for GG3-indicated regions(indicated by yellow arrows). *Labels*: Non-Cancer, GG3, GG4, GG5.

focused on low-grade (GG3) and high-grade (GG4+GG5) cancer. To prepare the tiles for training and evaluation, we applied the tissue detection and tile extraction techniques described in CLAM [26], while Mixed Supervision relied on SegGINI for data preparation. For slide-level classification models [3, 34], which depend on pre-trained embeddings, the extracted tiles are transformed into latent spaces using different models: ResNet50 pre-trained on ImageNet (a), ResNet50 pre-trained on TCGA and TULIP with MoCoV2 [9, 13, 20] (b), and ViT-small pre-trained on TCGA and TULIP with DINO [7, 20] (c). The main paper includes results for (c), which demonstrated the most superior performance regarding feature representation and cancer grading compared to others. Please refer to the Supplemental Document for (a) and (b).

On LDM's conditions. Global labels, such as Tile Labels, provide weak information, causing Latent Diffusion Models (LDMs) to predominantly learn a standard pattern associated with the tile label while ignoring other patterns present within the training tile. Consequently, it becomes challenging for LDMs to generate admixture of Gleason Grades (GGs). Furthermore, combining Tile and Slide Labels is not a logically sound approach as they are independent variables for tile synthesis; however, we do present results from LDMs conditioned by this combination in Supplemental Document. To overcome this problem, we lever-

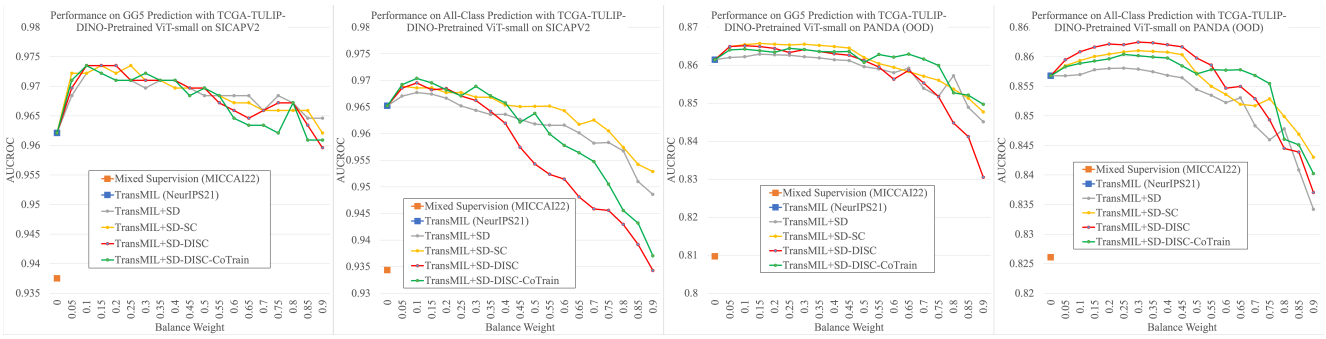

Figure 6. A quantitative comparison among TransMIL [34], Mixed Supervision [3], and TransMIL jointly trained with tiles generated by our models (discussed in Section 2.3) with a balance weight $\lambda \in [0.0, 0.9]$ in AUCROC. The feature representation extractor used is ViT-small (patch of 16) pre-trained on histopathology images with DINO [7, 20]. All models are trained on the SICAPv2 [35] and evaluated on both in-distribution SICAPv2 and Out-Of-Distribution (OOD) PANDA [5]. Our generated data consistently improves cancer grading performance with higher AUCROC. Please check our Supplemental Document for more results including the feature representation extractors ResNet50 pre-trained on ImageNet and histopathology images with MoCov2 [9, 13, 20].

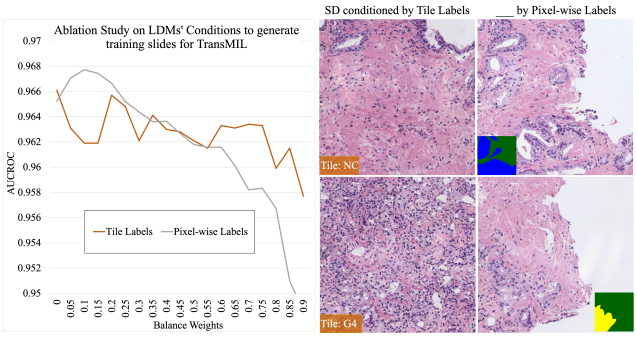

Figure 7. Ablation Study investigating the impact of Latent Diffusion Models' (LDMs) conditions including Tile Labels and our Pixel-wise Labels on enhancing TransMIL's performance (*left*) in AUCROC and qualitative evaluation (*right*). TransMIL utilizes feature representations pre-trained on histopathology images with ViT and DINO [20]. Consequently, LDMs conditioned with Pixel-wise Labels effectively allow a mix of Gleason Grades (GGs) in the tiles. Conversely, LDMs conditioned with Tile Labels tend to generate a single pattern per tile. Quantitatively, TransMIL trained on tiles conditioned by Pixel-wise Labels achieves the best performance with a BW of 0.1.

age pixel-wise labels available within SICAPv2 [35] and propose an efficient sampling technique to automatically generate tile sets without requiring further user annotation. Consequently, tiles generated with pixel-wise labels contain the more anticipated patterns enriched with pixel-level information. Training TransMIL on such tiles yields the most optimal cancer grading performance, outperforming the two other conditions, as shown in Figure 7. Please refer to Supplementary Document for layouts as generation guidance.

On improving the accuracy for histopathology image synthesis. In Figure 5, Latent Diffusion Models (LDMs) [30], known as Stable Diffusion (SD), with pixel-wise labels successfully generate high-fidelity tiles (second column) that closely resemble the actual tiles (first column). Nevertheless, when conditioned by pixel-wise multiple-GG-guided masks, SD tends to generate incorrect patterns in certain regions. For instance, it fails to generate any glands in the GG3-indicated region (second and last rows) and produces fused glands representing GG4 in the Non-Cancer-indicated region (third row). To address these issues and enhance the accuracy of Non-Cancer and GG patterns, we introduce Separated Conditions (SC) to generate distinct latent features, denoted as SD-SC (third column). However, generating K latent representations from K label masks significantly increases the computational cost. To mitigate this challenge, we propose Self-Distillation from Separated Conditions (DISC) and fine-tune the well-trained SD using DISC, denoted as SD-DISC. Additionally, we train SD-DISC with real tiles to maintain realism, wherein training errors from real and synthesized tiles are averaged and jointly optimized, denoted as SD-DISC-CoTrain. As a result, SD-DISC can effectively mimic the latent features obtained from SD-SC, providing accurate patterns similar to SD-SC (fourth column). Nonetheless, these generated tiles occasionally deviate from realism, as observed in the third row. To address this limitation, we further train SD-DISC with real data, bringing the generated tiles closer to SD in terms of realism (last column). More results can be found in Supplementary Document.

On improving pixel-level prostate cancer grading. Our study aims to enhance the performance of pixel-level prostate cancer grading models by incorporating additional views of training tiles. To validate the effectiveness of our approach, we conduct both quantitative and qualitative comparisons with the baseline Carcino-Net [25] and Carcino-Net trained on tiles generated by our ablation models (as detailed in Section 2.3). Our test sets comprise

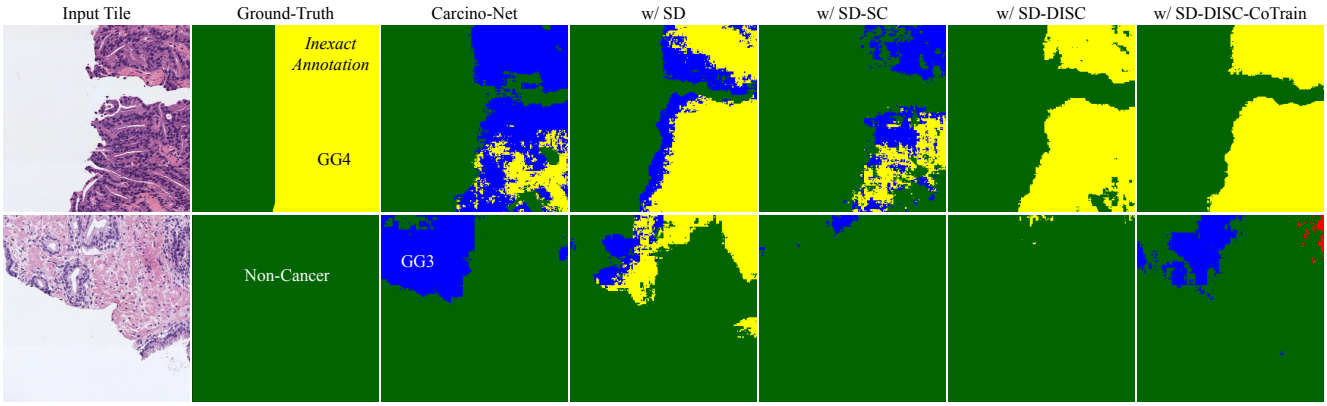

Figure 8. Qualitative comparison between Carcino-Net [25] and itself trained with our techniques discussed in Section 2.3.

2,100 tiles from the in-distribution SICAPv2 dataset [35] with Gleason Grade (GG) noisy pixel-wise annotations and 2,200 tiles from the out-of-distribution LAPC dataset [23] for low-grade (GG3) and high-grade (GG4+GG5). The ground-truth annotations in the SICAPv2 dataset are inexact and incomplete, usually mislabeling Non-Cancer patterns such as background and stroma as GGs and providing incomplete annotations for GGs. These inaccuracies reduce the reliability of quantitative evaluations. Models trained on this imprecise ground-truth often struggle with misclassifying Non-Cancer patterns. Nevertheless, we still report on the cancer grading accuracy for positive predictions of Gleason Grades, omitting Non-Cancer label and focusing on pixel-level precision in Supplementary Document. As a result, Carcino-Net trained with SD-SC, SD-DISC, and SD-DISC-CoTrain effectively segments out Non-Cancer patterns thanks to generated training data with accurate annotations. In contrast, other models tend to misclassify Non-Cancer as Gleason patterns, as shown in Fig. 8.

On improving slide-level prostate cancer grading. In this section, we assess the performance of previous works including TransMIL [34], its enhanced version Mixed Supervision [3], and TransMIL models jointly trained with tiles generated by our ablation models such as SD, SD-SC, SD-DISC, and SD-DISC-CoTrain. For all-class prediction, the baseline TransMIL model achieves an AUCROC of [**96.52%**, **85.68%**] on [**in-distribution SICAPv2**, **out-of-distribution PANDA**]. Meanwhile, tiles generated by SD yield improvements, resulting in [**96.77%**, **85.80%**] AUCROC (λ=0.1, λ=0.25). However, SD occasionally produces inaccurate patterns in specified regions, affecting the precision of synthetic training data. Addressing this, SD-SC is introduced and attains even better outcomes with AUCROC of [**96.89%**, **86.01%**] (λ=0.05, λ=0.3). SD-DISC-CoTrain, fine-tuned on in-distribution SICAPv2 while distilling from SD-SC, achieves top performance with an AUCROC of **97.04%** (λ=0.1) on SICAPv2. SD-DISC provides more generalized training tiles for out-of-distribution PANDA, with TransMIL+SD-DISC achieving top performance with an AUCROC of **86.25%** (λ=0.3). Additionally, our generative models improve performance for rare cases like GG5, with AUCROC improvements of up to [**+1.14%**, **+0.57%**]. Full and additional results on TransMIL enhancements with feature representations from two other extractors can be found in the Supplementary Document.

4. Conclusion

Latent Diffusion Models (LDMs) [30], also known as Stable Diffusion (SD), have demonstrated their potential in augmenting histopathology image tiles for training cancer grading models. In this study, we trained LDMs conditioned by human-annotated masks with multiple Gleason Grades (GGs). Furthermore, we introduced SD with Separated Conditions (SD-SC), which generates distinct latent features conditioned by separated conditions, to enhance the accuracy of generating patterns indicated by the complex GG-guided masks. However, SD-SC is associated with an increase in processing time. To mitigate this computational cost while maintaining performance, we proposed SD with Self-Distillation from Separated Conditions (DISC), allowing the SD model to mimic the latent features of SD-SC and generate improved GG patterns. As a result, our LDMs with DISC can produce higher-confidence patterns for guided masks, as in Figure 5. Additionally, when using our augmented data, pixel-level and slide-level cancer grading models such as CarcinoNet [25] and TransMIL [34] demonstrate improved performance compared to their baselines, particularly in the challenging GG5 cases. Our approach also surpasses the advanced Mixed Supervision [3] on both in-distribution and out-of-distribution data. In conclusion, our proposed LDMs with DISC offer a more accurate and effective approach for histopathology image augmentation, leading to improved cancer grading performance across different datasets and challenging GG categories.

5. Acknowledgements

We acknowledge the generous support from the Department of Defense Prostate Cancer Program Population Science Award (grant number W81XWH-21-0725); and also, the VA Merit Award (grant number 1 I01 CX002622-01). We also thank Dr. Akadiusz Gertych for the dataset from Cedars-Sinai Hospital in Los Angeles.

6. Compliance with Ethical Standards

LAPC images and annotations are available through a Material Transfer Agreement with Cedars-Sinai Hospital. The human subject data associated with the SICAPv2 dataset has been publicly released by [35] under the Creative Commons Attribution 4.0 International license[1].

References

[1] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018. 2

[2] Valentin Anklin, Pushpak Pati, Guillaume Jaume, Behzad Bozorgtabar, Antonio Foncubierta-Rodriguez, Jean-Philippe Thiran, Mathilde Sibony, Maria Gabrani, and Orcun Goksel. Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 636–646. Springer, 2021. 5

[3] Hao Bian, Zhuchen Shao, Yang Chen, Yifeng Wang, Haoqian Wang, Jian Zhang, and Yongbing Zhang. Multiple instance learning with mixed supervision in gleason grading. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 204–213. Springer, 2022. 5, 6, 7, 8

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 3, 5, 7

[6] Sujata Butte, Haotian Wang, Aleksandar Vakanski, and Min Xian. Enhanced sharp-gan for histopathology image synthesis. *arXiv preprint arXiv:2301.10187*, 2023. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 6, 7

[8] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. 2

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6, 7

[10] Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data*, 10(1):231, 2023. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised 824 visual representation learning. *arXiv1911. 05722 [cs]*, 825, 2019. 6, 7

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4

[16] Man M Ho and Jinjia Zhou. Deep photo scan: Semi-supervised learning for dealing with the real-world degradation in smartphone photo scanning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1880–1889, 2022. 2, 4

[17] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2019. 2

[18] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 3

[19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[20] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023. 6, 7

[1] https://data.mendeley.com/datasets/9xxm58dvs3/1

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[23] Wenyuan Li, Jiayun Li, Karthik V Sarma, King Chung Ho, Shiwen Shen, Beatrice S Knudsen, Arkadiusz Gertych, and Corey W Arnold. Path r-cnn for prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 38(4):945–954, 2018. 3, 5, 8

[24] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 648–663. Springer, 2020. 2

[25] Avinash Lokhande, Saikiran Bonthu, and Nitin Singhal. Carcino-net: A deep learning framework for automated gleason grading of prostate biopsies. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1380–1383. IEEE, 2020. 2, 3, 4, 5, 7, 8

[26] 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. 6

[27] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2

[28] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023. 1, 2

[29] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 3

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[33] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[34] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 2, 3, 4, 5, 6, 7, 8

[35] Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195: 105637, 2020. 3, 4, 5, 7, 8, 9

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[38] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022. 2

[39] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[41] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018. 2

[42] Mingkuan Yuan and Yuxin Peng. Ckd: Cross-task knowledge distillation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 22(8):1955–1968, 2019. 2

[43] Devora Zack. *Singletasking: Get More Done—One Thing at a Time*. Berrett-Koehler Publishers, 2015. 2

[44] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 2

[45] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2, 4