https://doi.org/10.31891/csit-2025-2-1
UDC 004.932: 616-006.04: 004.822

KUZMIN Sergii
Lviv Polytechnic National University
BEREZSKY Oleh
West Ukrainian National University

ANALYSIS OF DIFFUSION MODELS AND BIOMEDICAL IMAGE GENERATION TOOLS

Deep-learning pathology thrives on data, yet truly large, well-annotated histopathology archives remain rare. Synthetic slides created by text-to-image diffusion models promise relief, yet their usefulness depends on how faithfully they mimic real micro-architecture. We therefore set out to identify the fine-tuning strategy that allows Stable Diffusion 1.5 to reproduce the complex textures and staining patterns of colon tissue when only 664 patches per class (normal mucosa, serrated lesion, adenocarcinoma, adenoma) are available.

To that end, we fine-tuned four widely adopted methods - LoRA, DreamBooth, HyperNetwork and Textual Inversion - under identical computational and data constraints. We then evaluated the resulting images with the standard triad of Fréchet Inception Distance (FID), Precision and Recall. Our measurements reveal a clear hierarchy: HyperNetwork attains the lowest FID (≈ 77 on adenocarcinoma) while preserving high Precision and Recall, indicating sharp, diverse and anatomically coherent output. DreamBooth trails by a narrow margin, whereas Textual Inversion performs poorly (FID > 158; low Recall), confirming that embedding-only adaptation struggles with subtle glandular morphology.

While these metrics guide model selection, we recognise that quantitative scores alone cannot guarantee clinical adequacy. In our manual inspection we observed occasional artefacts - patchy stromal backgrounds or mild nuclear swelling - that slipped past automated checks, underscoring the need for medical professional review. Additionally we also note broader issues with the research of such type. First, generic computer-vision metrics overlook domain-specific cues; future work should develop histology-aware criteria that reward cell-level realism and diagnostic salience. Second, the community lacks guidance on adjusting diffusion hyper-parameters - such as LoRA rank or the layer mix in HyperNetwork - when training data remain below one thousand images per class. We contend that closing these gaps will convert synthetic slides from an intriguing demonstration into a dependable resource for medical education, quality assurance and algorithm development.

Keywords: diffusion models, stable diffusion, histopathology, image generation, medical imaging, fine-tuning methods.

КУЗЬМІН Сергій
Національний університет «Львівська політехніка»
БЕРЕЗЬКИЙ Олег
Західноукраїнський національний університет

АНАЛІЗ ДИФУЗІЙНИХ МОДЕЛЕЙ ТА ЗАСОБІВ ГЕНЕРУВАННЯ БІОМЕДИЧНИХ ЗОБРАЖЕНЬ

Патологічна діагностика, що ґрунтується на методах глибокого навчання, потребує значних масивів даних, однак справді великі й детально анотовані архіви гістопатологічних зображень залишаються рідкістю. Синтетичні слайди, створені дифузійними моделями типу «текст-до-зображення», можуть частково розв'язати цю проблему, проте їхня корисність залежить від точності відтворення реальної мікроархітектури тканин. Тому ми поставили за мету визначити, який метод тонкого налаштування дає змогу моделі Stable Diffusion 1.5 найкраще відтворювати складні текстури та забарвлення тканин товстої кишки, маючи лише по 664 патчі для кожного класу (нормальна слизова, зубчасте ураження, аденокарцинома, аденома).

Для цього ми донавчили модель чотирма поширеними підходами - LoRA, DreamBooth, HyperNetwork і Textual Inversion - за однакових обмежень обчислювальних ресурсів та даних. Далі якість згенерованих зображень оцінювалась стандартною тріадою метрик: Fréchet Inception Distance (FID), Precision і Recall. Отримані результати утворили чітку ієрархію: HyperNetwork забезпечує найнижчий FID (≈ 77 для аденокарциноми) при високих значеннях Precision і Recall, тобто формує чіткі, різноманітні й морфологічно узгоджені слайди. DreamBooth відстає лише незначно, тоді як Textual Inversion демонструє найгірші показники (FID > 158; низький Recall), що підтверджує обмеження підходу, заснованого лише на текстовому ембедингу.

Попри інформативність наведених метрик, одних лише кількісних оцінок недостатньо для гарантування клінічної придатності. Під час мануального огляду було виявлено окремі артефакти - плямистий фон строми чи помірне збільшення ядер - які автоматичні показники пропустили, що підкреслює необхідність експертної оцінки медичним працівником. Додатково нами також зазначено ширшу проблематику подібних досліджень. По-перше, універсальні метрики у комп'ютерному баченні не враховують галузеві нюанси; у майбутніх роботах слід розробити критерії, що відображають реалізм на клітинному рівні та діагностичну значущість. По-друге, бракує рекомендацій щодо налаштування гіперпараметрів дифузійних моделей - зокрема рангу LoRA чи комбінації шарів у HyperNetwork - коли навчальний набір налічує менше тисячі зображень на клас. Вирішення цих завдань, на нашу думку, перетворить синтетичні слайди з перспективної демонстрації на надійний ресурс для медичної освіти, контролю якості та розробки алгоритмів.

Ключові слова: дифузійні моделі, Stable Diffusion, гістопатологія, генерація зображень, біомедичні зображення, методи тонкого налаштування

Introduction

Modern medical diagnostics increasingly relies on the accurate classification of biomedical images, a task typically performed by convolutional neural networks (CNNs) [1]. However, these data-driven methods face

challenges due to the scarcity of large, well-annotated datasets in medical domains. Early generative methods, most notably Generative Adversarial Networks (GANs), attempted to address this limitation by offering synthetic data augmentation [2–4]. While GANs have contributed to performance gains across various imaging tasks, they often suffer from issues such as mode collapse and training instability. Recent advances in diffusion models [5–6], which operate without adversarial objectives, have shown superior stability and fidelity, making them particularly well-suited for complex biological structures. This is especially relevant in histopathology, where diffusion-driven approaches have already demonstrated promising results in generating synthetic images that closely match real tissues [7–8]. Yet the field lacks clear guidance on which fine-tuning strategy yields the most realistic and diagnostically informative histopathology images under tight data and computing budgets. By systematically benchmarking LoRA, DreamBooth, HyperNetwork and Textual Inversion on the same colon dataset, the present study aims to fill this gap and establish a data-efficient baseline for future medical-image augmentation efforts.

The object of the research – the generation of realistic histopathological images via diffusion-based generative modeling.

The subject of the research – the application of four specific fine-tuning strategies (LoRA, DreamBooth, HyperNetwork, Textual Inversion) to a pre-trained Stable Diffusion model for producing synthetic histopathology image data.

The purpose of the research – to identify which fine-tuning strategy for Stable Diffusion - LoRA, DreamBooth, HyperNetwork, or Textual Inversion - achieves the greatest image fidelity and diagnostic relevance when synthesizing histopathology images from limited single-organ data, thereby establishing a data-efficient benchmark for medical image augmentation.

To achieve this purpose, the following main research objectives are identified:

1. **Analyze** the performance of LoRA, DreamBooth, HyperNetwork, and Textual Inversion on Stable Diffusion with respect to histopathological image generation.
2. **Evaluate** the generated results using quantitative measures (FID, Precision, Recall) and discuss their practical relevance in medical and diagnostic workflows.

Related works

Diffusion models have emerged as a powerful alternative for high-fidelity image generation. Denoising Diffusion Probabilistic Models (DDPMs) introduced by Ho et al. (2020) marked a breakthrough, achieving remarkably realistic samples by learning to reverse a gradual noising process [9]. Subsequent improvements – including architectural optimizations and faster samplers – demonstrated that diffusion models can even surpass GAN benchmarks on natural image synthesis [10]. One key advantage of diffusion models is their training paradigm, which avoids adversarial minimax games and thus sidesteps many failure modes of GANs (e.g. collapse to limited modes), yielding more stable convergence and better coverage of the target distribution [7]. These models generate images by iteratively denoising random noise, a process that is computationally intensive but highly effective at capturing complex data distributions.

Diffusion models are rapidly gaining traction in biomedical imaging research. For instance, Pinaya *et al.* (2022) applied latent diffusion to brain MRI scans, demonstrating that synthetic neuroimaging data can closely mimic real scans [11]. Other works have leveraged diffusion for cross-modality translation (e.g. CT to MRI) using adversarial diffusion frameworks [12, 13], for accelerated MRI reconstruction via score-based diffusion sampling [14], and even for generating diverse segmentation masks by treating segmentation as a conditional generation task [15, 16]. In computational pathology, diffusion approaches are beginning to outperform prior GAN-based methods. Niehues *et al.* (2024) showed that a Stable Diffusion model fine-tuned on histology patches could generate tissue images which, when used to augment training data, boosted a classifier's accuracy by 4% [17].

Recent advancements in fine-tuning diffusion models have significantly enhanced their applicability in medical image synthesis. De Wilde et al. employed *Textual Inversion* to adapt pre-trained Stable Diffusion models for medical imaging modalities, demonstrating that training text embeddings on small datasets can produce diagnostically accurate images [18]. Peter et al. integrated fine-tuned Stable Diffusion with *DreamBooth* and *Low-Rank Adaptation (LoRA)* techniques, achieving high-fidelity medical image generation and highlighting the superior performance of Stable Diffusion in producing diverse, high-quality images [19]. Mao et al. introduced *SeLoRA*, a self-expanding LoRA module that dynamically adjusts its rank across layers during training, enhancing synthesis quality in medical imaging tasks [20]. Additionally, Ruiz et al. proposed HyperDreamBooth, a *HyperNetwork* capable of efficiently generating personalized weights from a single image, facilitating rapid personalization of text-to-image diffusion models with high subject fidelity [21]. These studies collectively showcase the efficacy of various fine-tuning methods in adapting diffusion models for medical image synthesis.

Materials and methods of research

This study examines the methods used to fine-tune the Stable Diffusion model for generating high-quality histopathology images. Specifically, we analyze techniques such as LoRA, DreamBooth, HyperNetwork, and Textual Inversion – each chosen for their ability to adapt pre-trained model to produce consistent, context-aware biomedical images. These fine-tuning methods were selected due to their widespread adoption and their balanced

trade-off between output quality and training efficiency, making them well-suited for scenarios where computational resources are limited and adaptation to a narrow domain, modality, or disease is required.

All the training conducted in this study is carried out utilizing the Stable Diffusion 1.5 pipeline with an image size of 512 × 512. Stable Diffusion 1.5 is a latent text-to-image diffusion model based on the architecture shown in Figure 1.

A stable diffusion model's architecture consists of a U-Net, a symmetric architecture with input and output of the same spatial size. The input image is first down-sampled and then up-sampled until reaching its initial size. U-Net consists of Wide ResNet blocks, group normalization, and self-attention blocks. The diffusion timestep t is specified by adding a sinusoidal position embedding into each residual block. The diffusion process is divided into two, namely forward and backward diffusion processes [23]. The equation below shows the forward process:

$$dx = f(x,t)dt = g(t)dw \tag{1}$$

Below is a backward diffusion process that reverses the time

$$dx = \left[f(x,t) - g(t)^2 \nabla_x \log p(x,t) \right] dt + g(t)dw \tag{2}$$

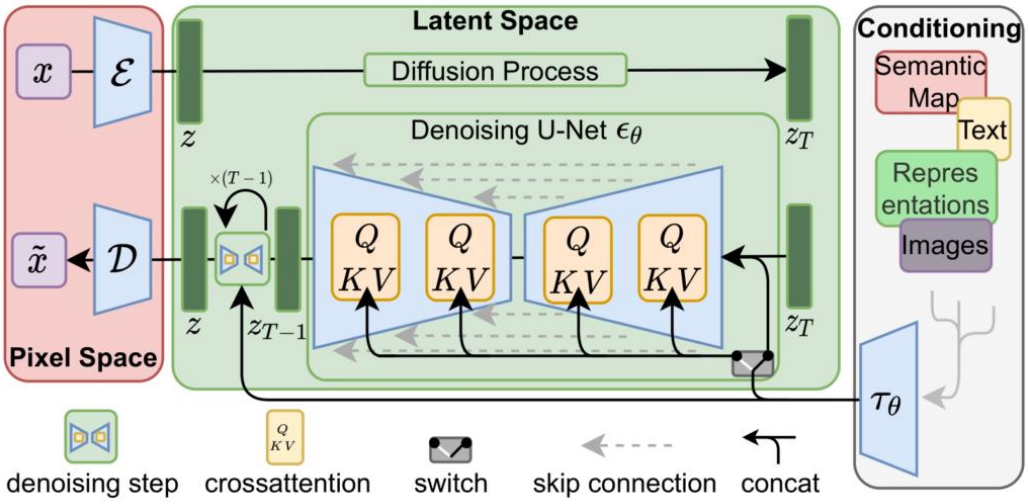

Fig. 1. The architecture of a latent diffusion model in a Stable Diffusion 1.5 pipeline [22]

Hence, the time-dependent score function $\nabla_x \log p(x,t)$ is known. Then, the diffusion process can be reversed. Training a diffusion model entails learning to denoise, for example, if a model can be scored using $f_\theta(x,t) \approx \nabla \log p(x,t)$ then denoising can be achieved by reversing the diffusion equation $x_t \to x_{t-1}$ [23]. The Score model $f_\theta : X \times [0,1] \to X$ can be referred to as a time-dependent vector field over X space. Hence, the Training objective is to first infer noise from a noised sample [23] as seen in the equation below:

$$\begin{aligned} &x \sim p(x), \varepsilon \sim \mathrm{N}(0,I), t \in [0,1] \\ &min \left\| \varepsilon + f_\theta \left(x + \sigma^t \varepsilon, t \right) \right\|_2^2 \end{aligned} \tag{3}$$

Furthermore, adding Gaussian noise ε to an image x with scale σ^t, helps the diffusion model to learn how to infer the noise σ. Another method of inferring noise is through what is known as conditional denoising. This method infers noise from a noised sample, based on a condition y:

$$\begin{aligned} &x, y \sim p(x,y), \varepsilon \sim \mathrm{N}(0,I), t \in [0,1] \\ &min \left\| \varepsilon - f_\theta \left(x + \sigma^t \varepsilon, y, t \right) \right\|_2^2 \end{aligned} \tag{4}$$

Hence, the conditional score model $f_\theta : X \times Y \times [0,1] \to X$ uses U-Net to model image-to-image mapping while modulating the U-Net with condition in the form of a text prompt.

To summarize, the Stable Diffusion architecture comprises of four (4) key components:

a) *VAE*: This algorithm compresses an image of 512 × 512 pixels into a 64 × 64 latent space.

b) *Forward and Reverse Diffusion*: Gaussian noise is progressively added in both forward and reverse diffusion until only random noise is present. It contributes to the uniqueness of the images.

c) *Noise Predictor (U-Net)*: Estimates and subtracts noise from the latent space to improve the visual output.

d) *Text Conditioning*: Stable Diffusion uses text prompts to introduce conditioning. Text prompts are analyzed by a CLIP tokenizer, which embeds them into a 768-value vector and uses a text transformer to direct the U-Net noise predictor.

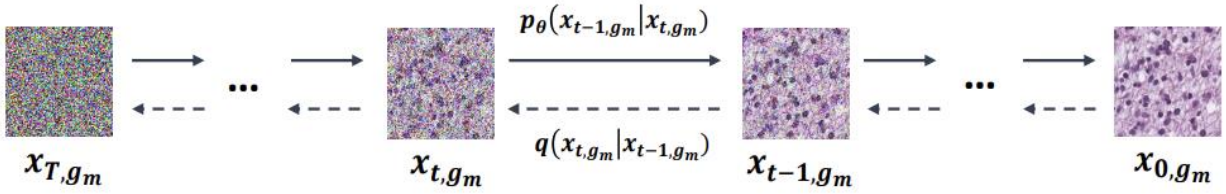

Fig. 2. The illustration of the forward (0 to T) and reverse (T to 0) diffusion process [7]

LoRA [24] is an approach that accelerates the finetuning of large models and is first proposed for language model adaptation. Instead of retraining all model parameters, LoRA adds pairs of rank-decomposition matrices and optimizes only these newly introduced weights. By limiting the trainable parameters and keeping the original weights frozen, LoRA is less likely to cause catastrophic forgetting. Concretely, the rank-decomposition matrices serve as the residual of the pre-trained model weights $W \in \mathrm{R}^{m\times n}$. The new model weight with LoRA is

$$W' = W + \Delta W = W + AB^T, \tag{5}$$

where $A \in \mathrm{R}^{m\times r}, B \in \mathrm{R}^{n\times r}$ are a pair of rank-decomposition matrices, r is a hyper-parameter, which is referred to as the rank of LoRA layers. In practice, LoRA is only applied to attention layers, further reducing the cost and storage for model fine-tuning.

DreamBooth method injects a specific visual subject into a text-to-image model with a unique token that refers to the subject [25]. It is achieved by fine-tuning the entire text-to-image diffusion model in two steps: (a) fine tuning the low-resolution text-to-image model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., "A photo of a [T] dog"), in parallel, a class-specific prior preservation loss is applied, which leverages the semantic prior that the model has on the class and encourages it to generate diverse instances belong to the subject's class by injecting the class name in the text prompt (e.g., "A photo of a dog"). (b) fine-tuning the super resolution components with pairs of low-resolution and high-resolution images taken from our input images set, which enables us to maintain high-fidelity to small details of the subject.

The HyperNetwork fine-tuning method offers an efficient and modular approach for adapting large-scale text-to-image diffusion models, such as Stable Diffusion, to new tasks or personalized concepts using a lightweight auxiliary network. Rather than updating the parameters of the primary generative model directly, this method leverages a secondary neural network – termed a *hypernetwork* – which dynamically generates or modulates weights for specific layers of the base model [21].

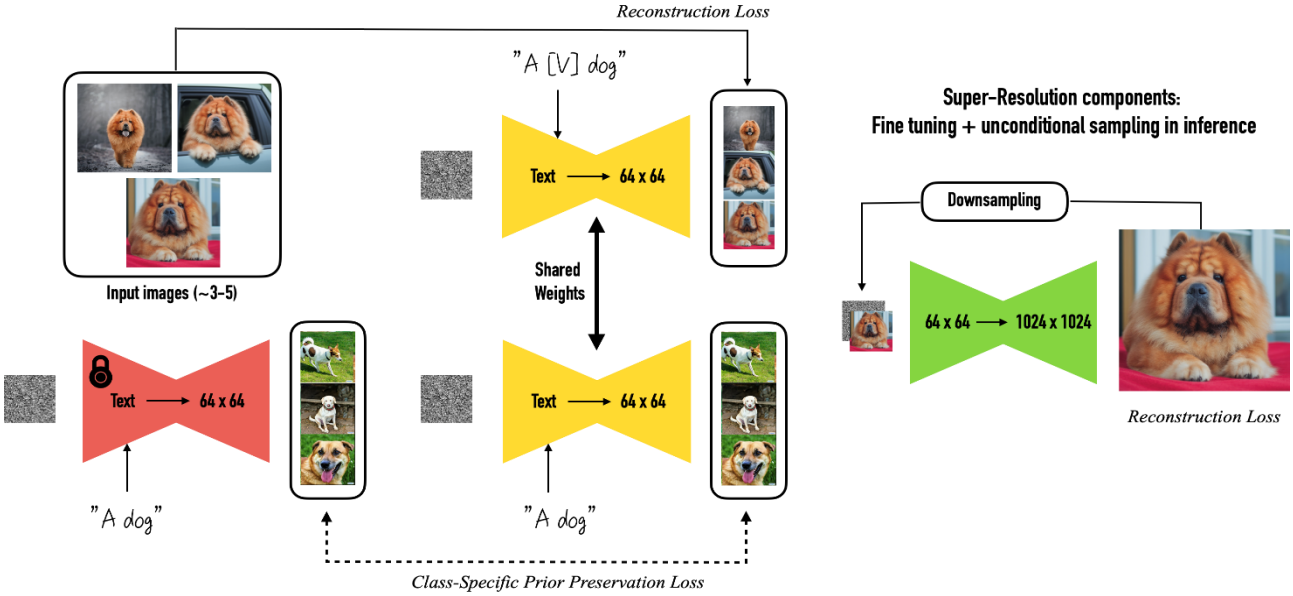

Fig. 3. DreamBooth fine-tuning of a diffusion model [25]

Formally, let $f(x,\theta)$ denote the pre-trained diffusion model parameterized by θ, and let ϕ denote the trainable parameters of the HyperNetwork $H(z,\phi)$. During inference, the effective weights of selected layers in the base model are adapted via:

$$\theta' = \theta + H(z,\phi), \tag{6}$$

where z is a learned task-specific or concept-specific embedding (e.g., representing a style or subject). The HyperNetwork is typically trained using a reconstruction or diffusion loss that minimizes the discrepancy between generated images and reference images, conditioned on textual prompts that may or may not include a concept placeholder token.

The primary benefit of this approach lies in its parameter efficiency and flexibility. By limiting the number of learnable parameters to the HyperNetwork alone – often orders of magnitude smaller than the full diffusion model – it becomes feasible to personalize or adapt the model with significantly reduced computational and memory overhead.

In practice, HyperNetwork is commonly attached to attention or feedforward projection layers within the U-Net or transformer blocks of the diffusion architecture. These modules learn to apply task-specific modulations, such as subject consistency, without degrading the general generation capabilities of the base model.

Stable Diffusion model can be conditioned on class labels, segmentation masks, or even on the output of a jointly trained text-embedding model. Let $c_\theta(y)$ be a model that maps a conditioning input y into a conditioning vector. The Latent Diffusion Model (LDM) loss is then given by:

$$L_{LDM} := \mathrm{E}_{z\sim \mathrm{E}(x),y,\varepsilon\sim \mathrm{N}(0,1),t}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t,t,c_\theta(y)\right)\right\|_2^2\right], \tag{7}$$

where E is the encoder that learns to map images x into a spatial latent code $z\sim \mathrm{E}(x)$, t is the time step, z_t is the latent noised to time t, ε is the unscaled noise sample, and ε_θ is the denoising network. In Stable Diffusion 1.5 model c_θ is realized through CLIP Text Encoder (specifically the ViT-L/14 variant).

Text encoder models begin with a text processing step (Figure 4). First, each word or sub-word in an input string is converted to a token, which is an index in some pre-defined dictionary. Each token is then linked to a unique embedding vector that can be retrieved through an index-based lookup. These embedding vectors are typically learned as part of the text encoder c_θ.

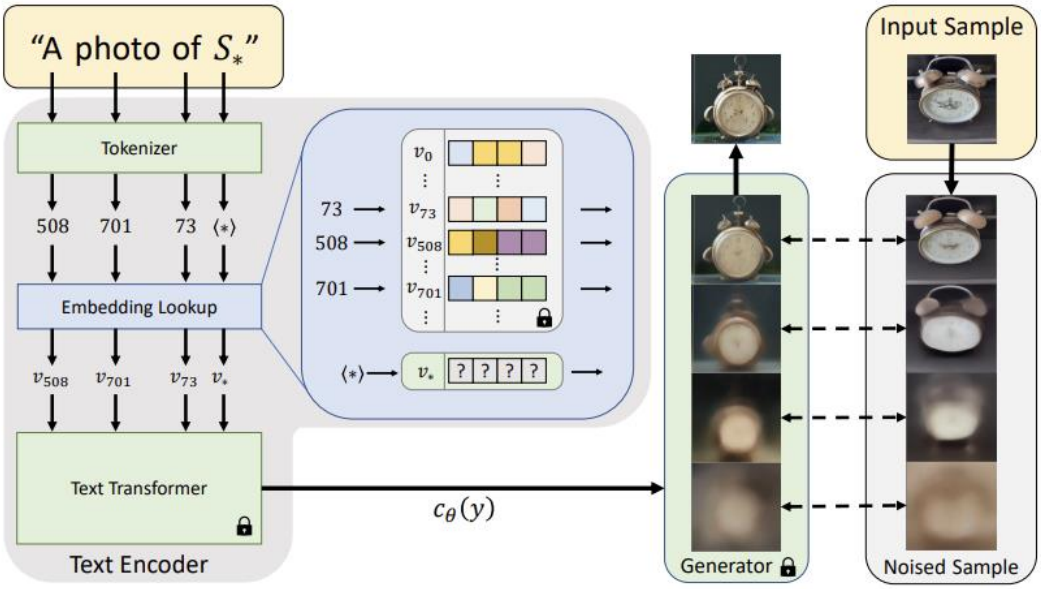

Fig. 4. Outline of the text-embedding and inversion process [26]

This embedding space is the target for textual inversion. Specifically, a placeholder string S_* is designated to represent the new concept. The method intervenes in the embedding process and replaces the vector associated with the tokenized string with a new, learned embedding v_*, in essence "injecting" the concept into model's

vocabulary. v_* is found through direct optimization, by minimizing the LDM loss (7) over images sampled from the small set that can be defined as:

$$v_* = \arg\min_{v} \mathrm{E}_{z\sim\mathrm{E}(x),y,\varepsilon\sim\mathrm{N}(0,1),t}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, c_\theta\left(y\right)\right)\right\|_2^2\right], \quad (8)$$

and is realized by re-using the same training scheme as the original LDM model, while keeping both c_θ and ε_θ.

Experiments

In this section, we conduct fine-tuning experiments on Stable Diffusion 1.5 model using proposed methods and perform quantitative quality assessment of the generated synthetic histopathology images.

We perform fine-tuning and evaluation on publicly available Chaoyang dataset, which contains colon slides from Chaoyang hospital with 512 × 512 patch size [27]. The dataset itself contains 1111 normal, 842 serrated, 1404 adenocarcinoma, 664 adenoma, and 705 normal, 321 serrated, 840 adenocarcinoma, 273 adenoma samples for training and testing sets, respectively. Figure 5 shows the sample patches.

For our experiments we decided to rebalance the dataset, so that the same number of images is used across different tissue types for both training and testing sets. 664 samples of each type for the training set and 273 samples of each type for the testing set were randomly selected from the base dataset.

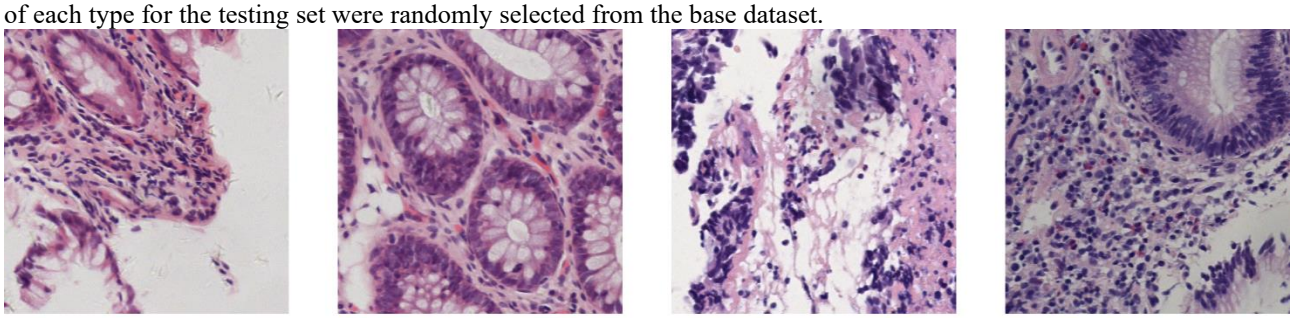

Fig. 5. Histological samples from Chaoyang dataset [27]

One of the main reasons for such rebalancing stems from the fact that Textual Inversion is typically limited to single-concept fine-tuning, while LoRA, DreamBooth, and HyperNetwork more naturally accommodate multi-concept fine-tuning within a single model. For fair comparisons the Stable Diffusion is fine-tuned with each method on a single concept (tissue type) at a time separately using the same dataset size, allowing direct comparisons across all methods, including Textual Inversion.

All the training is carried out utilizing the base Stable Diffusion 1.5 model (stabilityai/stable-diffusion-v1-5) on the local environment with Windows 11 Pro 64-bit operating system and the following hardware, which reflects a configuration with limited VRAM at disposal: Intel Core i9-13980HX, 32 GB DDR 5 RAM, and NVIDIA GeForce RTX 4090 Laptop GPU with 16 GB of GDDR6 VRAM.

LoRA and DreamBooth trainings are conducted using Kohya_SS v24.1.7 [28] – a software environment, which provides an accessible interface for users to fine-tune the model's parameters effectively. Textual Inversion training is executed using OneTrainer (commit fe59c2a) [29], which is the alternative tool for Stable Diffusion training. HyperNetwork training and image generation with all fine-tuned models is done using Stable Diffusion web UI v1.10.1 (AUTOMATIC1111) [30], an open-source, browser-based interface designed to facilitate the use of the Stable Diffusion model for generating images.

The training parameters for each fine-tuning method are presented below.

Table 1

Training parameters for LoRA and DreamBooth fine-tuning.

LoRA		**DreamBooth**	
Repeat Factor	8	Repeat Factor	2
Batch Size	8	Batch Size	2
Epoch	12	Epoch	12
Mixed Precision	fp16	Mixed Precision	fp16
Warmup steps	5%	Warmup steps	5%
Text Encoder Learning Rate	5e-5	Text Encoder Learning Rate	5e-5
Unet Learning Rate	7e-5	Unet Learning Rate	1e-5
Network Rank	8	Optimizer	AdamW 8-bit
Network Alpha	8	Seed	22
Optimizer	AdamW 8-bit		
Seed	22		

Table 2

Training parameters for Textual Inversion and HyperNetwork fine-tuning

Textual Inversion		**HyperNetwork**	
Repeat Factor	2	Batch Size	4
Batch Size	32	Base Learning Rate	5e-5
Epoch	12	Layer Structure	[1, 2, 4, 2, 1]
Mixed Precision	fp16	Activation Function	Linear
Warmup steps	5%	Training Steps	4500
Text Encoder Learning Rate	5e-4		
Optimizer	AdamW 8-bit		

Since the dataset does not contain annotations, we relied solely on class labels to condition the model during training. Typically, prompts are generated from a list of templates such as "a photo of a <subject>" or "a rendering of a <subject>". However, these formats are not necessarily applicable in a medical imaging context. Therefore, we used only the raw class label (e.g., "normal_tissue", "adenoma_tissue") as the training prompt.

The inference was performed using DPM++ 2M sampling method in combination with the Karras scheduler, which provides improved stability and high-quality outputs, especially at higher step counts. The number of sampling steps was set to 50, offering a good balance between image fidelity and generation time. A CFG (classifier-free guidance) scale of 7.5 was used to control the strength of prompt conditioning, promoting prompt adherence while maintaining some generative flexibility. All images were generated at a resolution of 512×512 pixels, consistent with the training setup. A qualitative comparison of the synthesized images is presented in the figures below.

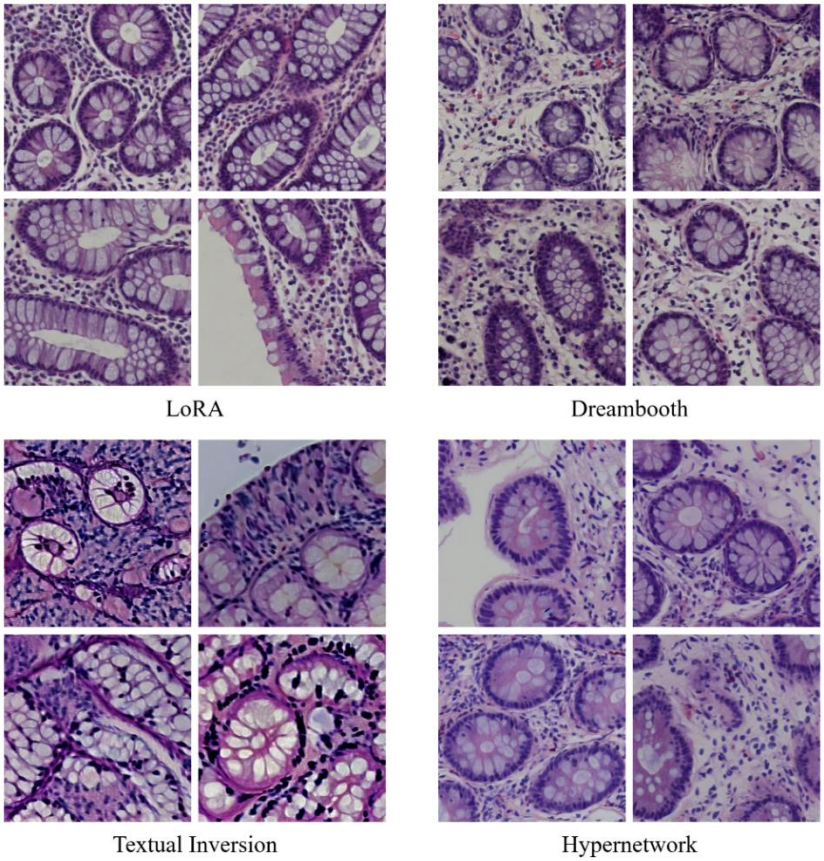

Fig. 6. Selection of generated patches of normal tissue

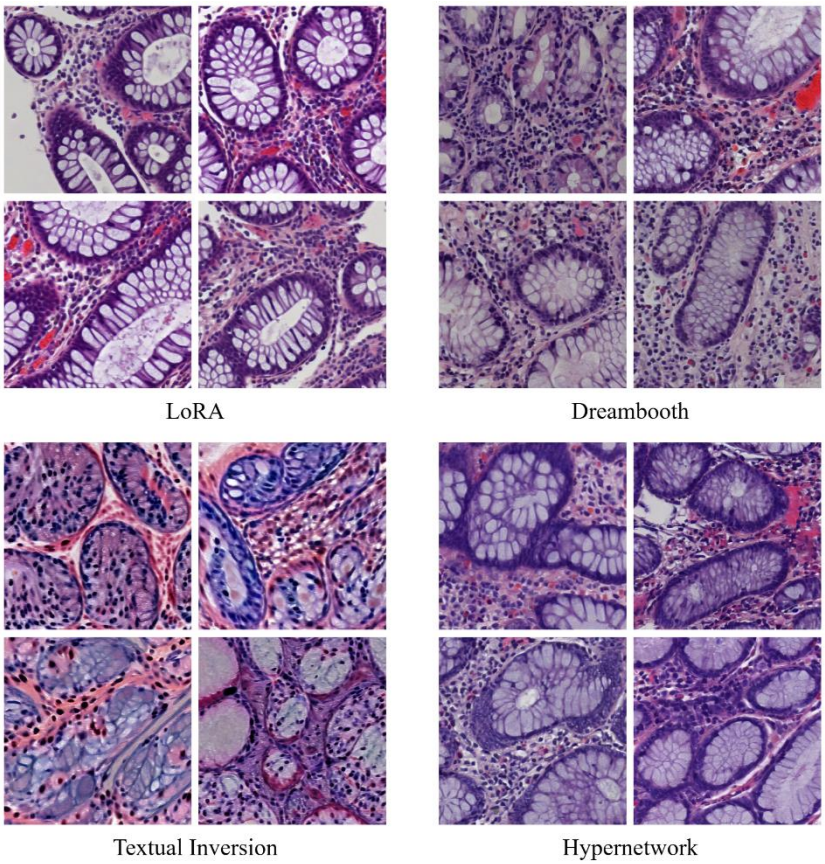

Fig. 7. Selection of generated patches of serrated tissue

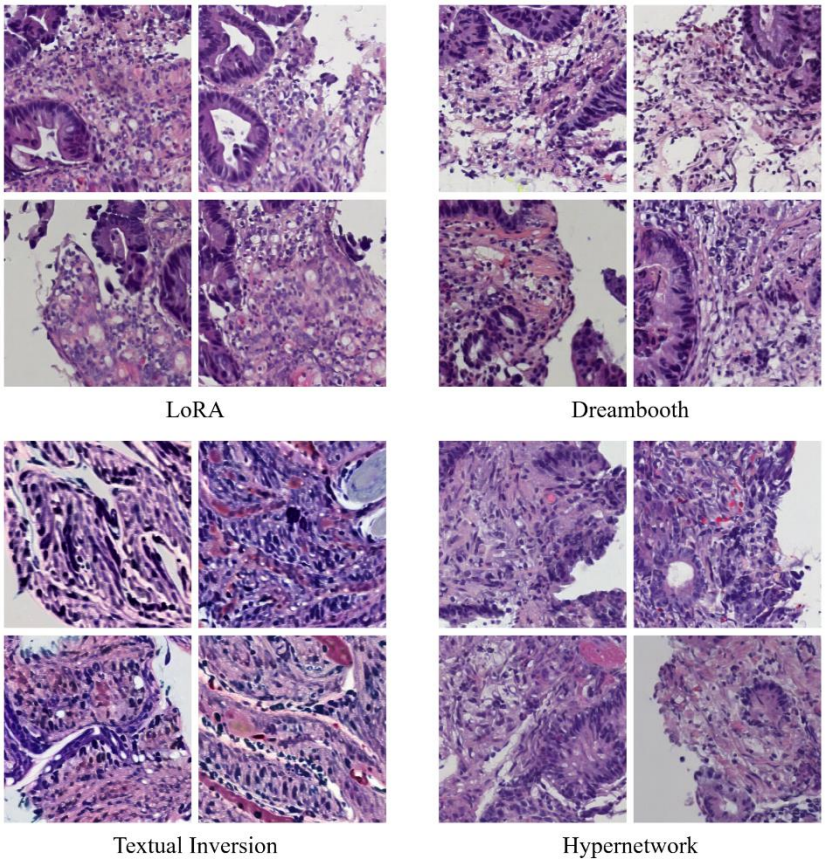

Fig. 8. Selection of generated patches of adenocarcinoma tissue

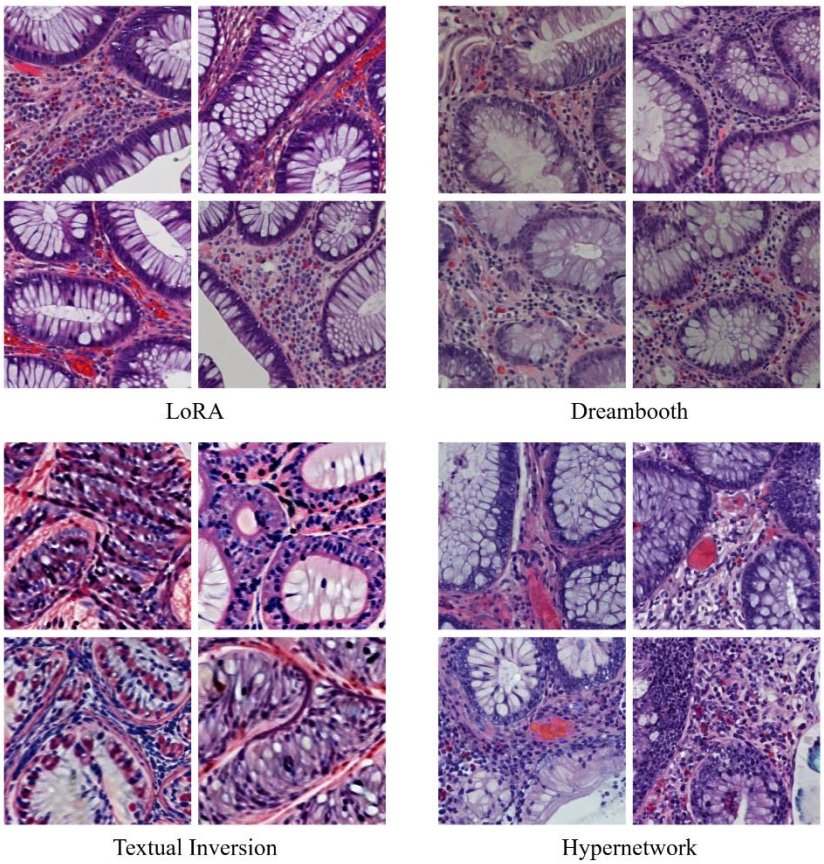

Fig. 9. Selection of generated patches of adenoma tissue

To quantitatively assess the quality of the synthesized images, we employed three widely used metrics in generative model evaluation: Fréchet Inception Distance (FID), Precision, and Recall. These metrics evaluate realism, fidelity, and diversity of generated samples by comparing them to real data in a learned feature space.

FID measures the distance between the feature distributions of real and generated images extracted from a pretrained Inception-V3 network. Specifically, it assumes both distributions are Gaussian and computes the Fréchet distance between them:

$$FID(X,Y) = \left\|\mu_x - \mu_y\right\|^2 + Tr(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{1/2}) \tag{9}$$

where μ_x, Σ_x and μ_y, Σ_y are the empirical means and covariances of the features from real samples X and generated samples Y, respectively. Lower FID scores indicate that the generated distribution is closer to the real data distribution, implying higher image fidelity.

Precision quantifies the fidelity of generated images by estimating the fraction of generated samples that lie within the support of the real data manifold. Formally, given a distance threshold ε, precision is computed as:

$$Precision = \frac{1}{|Y|}\sum_{y\in Y} I[\exists x \in X, \left\|f(x) - f(y)\right\| < \varepsilon] \tag{10}$$

where $f(\cdot)$ denotes feature embeddings, and $I[\cdot]$ is the indicator function. A high precision value indicates that generated images are realistic and similar to real samples.

Recall measures the diversity of generated images by evaluating how well the generative model covers the variety of real images:

$$Recall = \frac{1}{|X|}\sum_{x\in X} I[\exists y \in Y, \left\|f(x) - f(y)\right\| < \varepsilon] \tag{11}$$

High recall indicates that the model can generate a wide range of realistic samples that span the diversity present in the real dataset.

Together, these metrics provide a comprehensive view of generative model performance, balancing visual quality (FID, Precision) and variability (Recall). The calculated metrics are provided in Table 3 and Table 4.

Table 3

FID, Precision, and Recall metrics for images of normal and serrated tissue generated with four fine-tuning methods: LoRA, DreamBooth, Textual Inversion, and HyperNetwork.

	normal			**serrated**		
	FID	Precision	Recall	FID	Precision	Recall
LoRA	127,33	0,1209	0,6190	134,59	0,1026	0,4505
DreamBooth	**97,12**	**0,3443**	0,6850	127,06	0,2308	**0,7766**
Textual Inversion	158,58	0,2234	0,0256	171,59	0,2418	0,0733
HyperNetwork	101,22	0,1392	**0,7253**	**107,59**	**0,4579**	0,5934

Table 4

FID, Precision, and Recall metrics for images of adenocarcinoma and adenoma tissue generated with four fine-tuning methods: LoRA, DreamBooth, Textual Inversion, and HyperNetwork.

	adenocarcinoma			**adenoma**		
	FID	Precision	Recall	FID	Precision	Recall
LoRA	113,49	0,1245	0,3810	138,15	0,0623	0,2051
DreamBooth	122,14	**0,5531**	0,5128	121,61	0,1245	**0,7619**
Textual Inversion	164,03	0,2967	0,0659	147,33	0,2381	0,0440
HyperNetwork	**77,27**	0,3883	**0,5678**	**87,34**	**0,3223**	0,6557

Conclusions

In this study, we evaluated several prominent fine-tuning methods - LoRA, DreamBooth, Textual Inversion, and HyperNetwork - for synthesizing histopathological images using the Stable Diffusion 1.5 model. Quantitatively, HyperNetwork delivered the best image fidelity and diversity (FID ≈ 77 for adenocarcinoma, with high Precision and Recall), closely followed by DreamBooth. By contrast, Textual Inversion lagged far behind (FID > 158 and markedly lower Recall), confirming substantial performance gaps among the four fine-tuning methods. Although it remains possible that further parameter tuning could enhance Textual Inversion outputs to some degree, the inherent limitations of this technique - primarily due to its simplistic embedding-based approach - suggest that it may remain suboptimal relative to methods that more extensively modify model parameters, such as DreamBooth or HyperNetwork.

The Scientific novelty of the obtained research results – they present a first direct, in-depth comparison of four distinct fine-tuning methods (LoRA, DreamBooth, HyperNetwork, and Textual Inversion) in the context of large-scale diffusion modeling specifically for histopathological image generation, thereby elucidating the trade-offs among fidelity, diversity, and resource constraints.

The Practical significance of the research results – they offer a clear basis for selecting fine-tuning strategies that optimize synthetic data generation, enhance diagnostic model performance, and streamline resource allocation in medical imaging workflows.

Study limitations – several factors constrain the generalizability of the findings. First, the experiments rely on the Chaoyang colon-histopathology dataset, re-balanced to 664 training images per tissue class; this single-organ scope and modest sample size may not reflect the heterogeneity of broader histopathology collections. Second, all fine-tuning and inference were performed at a resolution of 512 × 512 pixels on hardware limited to 16 GB VRAM, so resolution-dependent artefacts or memory-intensive configurations remain unexplored. Third, image quality was assessed exclusively with generic computer-vision metrics (FID, Precision, Recall); while standard, these do not measure diagnostic appropriateness, making expert pathologist review indispensable. Finally, we evaluated the four fine-tuning strategies using specific set of training settings; adjustments to rank, learning-rate schedules, or regularization methods could shift the relative performance hierarchy observed here. These constraints should be kept in mind when interpreting the reported gains.

For future research, we recommend excluding the Textual Inversion approach and focusing analysis on methods better suited for multi-concept scenarios, particularly HyperNetwork and DreamBooth. These methods more naturally support training and generating images of various tissue types within a single model, making them practical for real-world medical applications.

Moreover, there is a clear need to explore and develop evaluation metrics specifically tailored to the medical imaging domain, especially histopathology. Such domain-specific metrics would provide a better alignment between quantitative evaluations and qualitative assessments conducted by medical experts, ultimately improving the relevance and interpretability of generated medical images.

Finally, while the general fine-tuning approaches used in this study are effective, achieving significant improvements in image synthesis quality may necessitate the development and application of fine-tuning techniques explicitly designed for histopathological imaging. Domain-specific fine-tuning methods have the potential to substantially enhance model performance, as evidenced by previous research in specialized medical contexts, indicating an important avenue for future methodological advancement.

References

1. S. Albawi, T. A. Mohammed, S. Al-Zawi. Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)* : Antalya, Turkey, 2017, pp. 1-6
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative Adversarial Networks. *Advances in Neural Information Processing Systems 27* : Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montreal, Quebec, Canada, 2014.
3. You-Bao Tang, Sooyoun Oh, Yu-Xing Tang, Jing Xiao, Ronald M. Summers. CT-Realistic Data Augmentation Using Generative Adversarial Network for Robust Lung Nodule Detection. *Proceedings Volume 10950, Medical Imaging 2019: Computer-Aided Diagnosis* : February 16-21, 2019, San Diego, California, United States, 2019.
4. D. Popescu, M. Deaconu, L. Ichim, G. Stamatescu. Retinal Blood Vessel Segmentation Using Pix2Pix GAN. *2021 29th Mediterranean Conference on Control and Automation (MED)* : PUGLIA, Italy, 2021, pp. 1173-1178
5. Dhariwal, P., Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems 34* : *Proceedings of the* 35th Conference on Neural Information Processing Systems (NeurIPS 2021), December 6-14, 2021, pp. 8780-8794.
6. Müller-Franzes, G., Niehues, J.M., Khader, F. et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Report* : vol. 13, 12098, Springer Science and Business Media LLC, 2023.
7. P. A. Moghadam et al. A Morphology Focused Diffusion Probabilistic Model for Synthesis of Histopathology Images. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* : Waikoloa, HI, USA, 2023, pp. 1999-2008
8. Rai, T. et al. Evaluating diffusion model generated synthetic histopathology image data against authentic digital pathology images. *Proceedings Volume 12933, Medical Imaging 2024: Digital and Computational Pathology* : SPIE Medical Imaging, February 18-23, 2024, San Diego, California, United States, 2024.
9. Ho, J., Jain, A., Abbeel, P. Denoising diffusion probabilistic models. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* : December 6-12, 2020, Curran Associates Inc., Red Hook, NY, USA, Article 574, pp. 6840–6851.
10. Nichol, A. Q., Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. *Proceedings of the 38th International Conference on Machine Learning (ICML)* : vol. 139, July 18-24, 2021, PMLR, pp. 8162-8171.
11. Pinaya, W.H.L. et al. Brain Imaging Generation with Latent Diffusion Models. *Deep Generative Models* : DGM4MICCAI 2022, Lecture Notes in Computer Science, vol 13609. Springer, Cham, pp. 117-126.
12. M. Özbey et al. Unsupervised Medical Image Translation with Adversarial Diffusion Models. *IEEE Transactions on Medical Imaging* : vol. 42, no. 12, December 2023, pp. 3524-3539.
13. Y. Luo et al. Target-Guided Diffusion Models for Unpaired Cross-Modality Medical Image Translation. *IEEE Journal of Biomedical and Health Informatics* : vol. 28, no. 7, July 2024, pp. 4062-4071.
14. Lyu, Q., Wang, G. Conversion Between CT and MRI Images Using Diffusion and Score-Matching Models. : September 2022 (Preprint. 10.48550/arXiv.2209.12104)
15. Hejrati, B., Banerjee, S., Glide-Hurst, C., Dong, M. Conditional Diffusion Model with Spatial Attention and Latent Embedding for Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* : Lecture Notes in Computer Science, vol 15009, 27th International Conference, Marrakesh, Morocco, October 6–10, 2024. Springer, Cham, pp 202–212.
16. Chen, T., Wang, C., Shan, H. BerDiff: Conditional Bernoulli Diffusion Model for Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* : Lecture Notes in Computer Science, vol 14223, 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023. Springer, Cham, pp 491–501.
17. Niehues, J. M., et al. Using histopathology latent diffusion models as privacy preserving dataset augmenters improves downstream classification performance. *Computers in Biology and Medicine* : vol. 175, June 2024, Article 108410.
18. de Wilde, et al. Medical Diffusion on a Budget: Textual Inversion for Medical Image Generation. *Medical Imaging with Deep Learning - MIDL 2024* : July 3, 2024, Paris France, 2024.
19. Peter, O. O. E., Rahman, M. M., Khalifa, F. Advancing AI-Powered Medical Image Synthesis: Insights from MedVQA-GI Challenge Using CLIP, Fine-Tuned Stable Diffusion, and Dream-Booth + LoRA. *CLEF 2024: Conference and Labs of the Evaluation Forum* : September 09–12, 2024, Grenoble, France.
20. Mao, Y., et al. SeLoRA: Self-Expanding Low-Rank Adaptation of Latent Diffusion Model for Medical Image Synthesis. *The Thirteenth International Conference on Learning Representations - ICLR 2025* : preprint, April 24, 2025, Singapore.
21. Ruiz, N., et al. HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models. *Proceedings of CVPR 2024 (IEEE/CVF Conference on Computer Vision and Pattern Recognition)* : June 2024, pp. 6527-6536.
22. Rombach, R., et al. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* : New Orleans, LA, USA, 2022, pp. 10674-10685.
23. Wang, B., Vastola, J. Stable diffusion. [Manuscript]. 2022. URL: https://scholar.harvard.edu/files/binxuw/files/stable_diffusion_a_tutorial.pdf
24. Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *The Tenth International Conference on Learning Representations - ICLR 2022* : Poster, Virtual, April 25, 2022.
25. Ruiz, N., et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* : Vancouver, BC, Canada, 2023, pp. 22500-22510.
26. Gal, R., et al. An Image is Worth One Word: Personalizing Text-to-Image Generation Using Textual Inversion. *The Eleventh International Conference on Learning Representations - ICLR 2023* : May 1, 2023.
27. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M. Hard Sample Aware Noise Robust Learning for Histopathology Image Classification. *IEEE Transactions on Medical Imaging* : vol. 41, no. 4, April 2022, pp. 881-894.
28. Bmaltais. (n.d.). bmaltais/kohya_ss. GitHub. URL: https://github.com/bmaltais/kohya_ss
29. Nerogar. (n.d.). Nerogar/OneTrainer. GitHub. URL: https://github.com/Nerogar/OneTrainer
30. AUTOMATIC1111. (n.d.). Stable Diffusion Web UI. GitHub. URL: https://github.com/AUTOMATIC1111/stable-diffusion-webui

Sergii Kuzmin **Сергій Кузьмін**	PhD Student at Department of Automated Control Systems. Lviv Polytechnic National University, Lviv, Ukraine, email: sergii.o.kuzmin@lpnu.ua; kuzminos22@gmail.com, https://orcid.org/0009-0001-7182-2883	аспірант кафедри "Автоматизованих систем управління", Національний університет «Львівська політехніка», Львів, Україна
Oleh Berezsky **Олег Березький**	Dr. Sci. (Eng), Professor, Head of Computer Engineering Department. West Ukrainian National University, Ternopil, Ukraine, email: ob@wunu.edu.ua; oleh.m.berezkyi@lpnu.ua, https://orcid.org/0000-0001-9931-4154, Scopus ID: 6505609877	доктор технічних наук, професор, завідувач кафедри Комп'ютерної інженерії. Західноукраїнський національний університет, вул. Львівська, Тернопіль, Україна