Efficient Self-Supervised Vision Transformers for Histopathology Image Retrieval

Marija Habijan[1], Petar Nakić[2], Irena Galić[1], Danijel Marinčić[3], Josip Samardžić[3], Aleksandra Pižurica[4]

[1] Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Osijek, Croatia
[2] University of Slavonski Brod, Slavonski Brod, Croatia
[3] General Hospital Dr. Josip Benčević Slavonski Brod, Slavonski Brod, Croatia
[4] Ghent University, Department of Telecommunications and Information Processing, TELIN-GAIM, Belgium

marija.habijan@ferit.hr

***Abstract*—Histopathology image retrieval is a crucial task in computational pathology, facilitating case-based reasoning and aiding in diagnosis and research. This work presents an efficient retrieval system based on Efficient Self-Supervised Vision Transformers (EsViT), which generates robust feature representations without labeled data. We fine-tune a pre-trained EsViT model on multiple histopathology datasets, including BracS, CRC and BATCH. The system extracts deep feature embeddings from query images and retrieves similar cases from an indexed feature database using k-nearest neighbors (k-NN) search. Experimental results show that EsViT-based retrieval significantly outperforms CNN-based approaches, achieving mAP@10 of 78.4% (BracS), 74.9% (CRC), and 79.1% (BATCH), with Precision@10 up to 81.2%. Additionally, EsViT reduces feature extraction time to 7.5 ms per image, making it faster and more scalable than CNN-based methods. Our findings demonstrate that self-supervised vision transformers enable accurate, scalable histopathology image retrieval, with applications in digital pathology and clinical decision support.**

***Index Terms*—histopathology image retrieval, self-supervised learning, vision transformers, EsViT, deep feature extraction, medical imaging.**

I. Introduction

Histopathology plays fundamental role in cancer diagnosis and prognosis, where microscopic tissue images are analyzed to enable pathologists to detect and classify abnormalities [1]. With the increasing availability of digitized whole-slide images (WSIs), computational methods for histopathology image retrieval have gained significant attention [2]. Efficient image retrieval systems can help find similar cases, supporting clinical decision-making, research, and educational purposes [3], [4].

Early content-based image retrieval (CBIR) systems for histopathology images relied on handcrafted feature extraction, using descriptors such as color histograms, texture features (e.g., Gabor filters, Local Binary Patterns), and shape-based representations. These methods typically employed Bag-of-Visual-Words (BoVW) models or Support Vector Machines (SVMs) to measure similarity between images. While computationally efficient, these techniques suffered from limited generalization due to their dependence on manually defined features, which struggled to adapt to the high variability of histopathology images.

The advent of deep learning significantly transformed histopathology image retrieval, shifting the focus toward automated feature extraction using convolutional neural networks (CNNs) [5]. Pre-trained architectures such as VGG, ResNet, and DenseNet have been widely adopted for feature embedding generation, enabling more robust image representations [6]. These deep embeddings are then indexed using k-nearest neighbor (k-NN) search or approximate nearest neighbor methods to facilitate efficient retrieval [7]. Further improvements have been achieved through fine-tuning CNNs on domain-specific datasets [8], ensuring that feature representations capture the intricate structures found in histopathological samples.

Beyond conventional CNN-based approaches, advanced deep learning techniques have been explored to enhance retrieval performance. Multi-scale CNNs and attention-based architectures have been introduced to improve feature discrimination by capturing both local and global tissue structures [9]. In addition, graph-based deep learning models have been proposed to represent spatial relationships between cellular structures [10], allowing for more biologically meaningful image retrieval. Generative adversarial networks (GANs) [11] have also been used for data augmentation and feature refinement, helping to mitigate the challenges posed by small, imbalanced datasets.

Despite these advancements, supervised deep learning models remain highly dependent on large annotated datasets, which are often costly and time-consuming to obtain. To address this challenge, self-supervised learning (SSL) has emerged as a promising alternative, enabling models to learn rich feature representations from unlabeled data. Techniques such as contrastive learning (e.g., SimCLR, MoCo) [12], [13], masked image modeling, and clustering-based pretraining have demonstrated strong performance in histopathology image analysis. By leveraging SSL, retrieval models can reduce reliance on labeled datasets while improving generalization to unseen cases. This shift toward self-supervised and transformer-based methods has the potential to further enhance histopathology

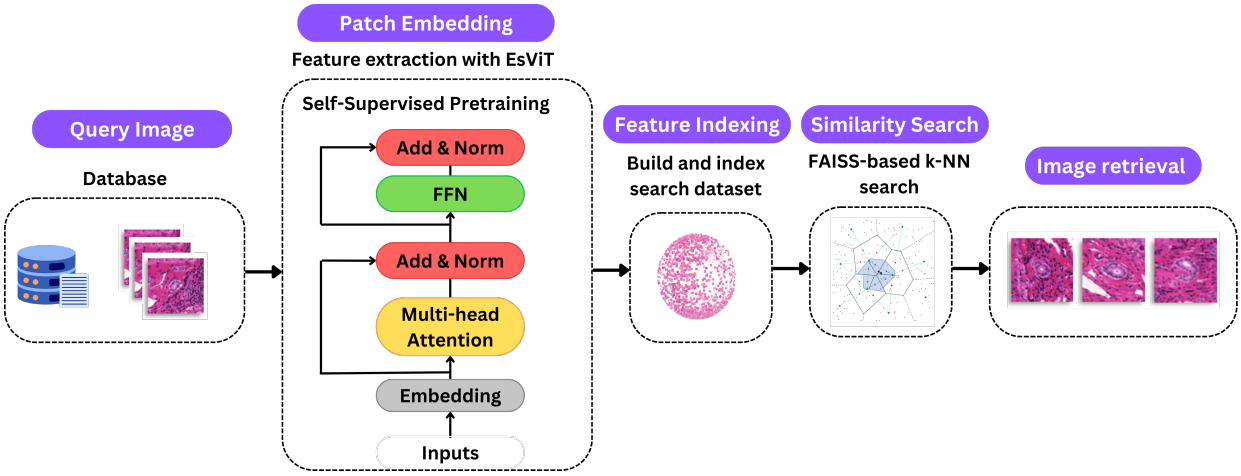

Fig. 1. Overview of the proposed EsViT-based histopathology image retrieval system. The query image is first tokenized into patches and processed through the EsViTs, to generate deep feature embeddings. These embeddings are then indexed using FAISS and stored in a structured database. During retrieval, a similarity search using k-nearest neighbors (k-NN) is performed in the feature space to identify the most relevant histopathology images. The retrieved images are ranked based on their similarity to the query image, supporting efficient and accurate case-based retrieval in digital pathology.

image retrieval, paving the way for more robust and scalable solutions [14].

In this work, we propose a histopathology image retrieval system based on Efficient Self-supervised Vision Transformers (EsViT) [15]. EsViT is a self-supervised vision transformer (ViT) [16] that efficiently learns visual representations by using h-based encoding, region matching, and hierarchical feature extraction. By fine-tuning a pre-trained EsViT model on histopathology datasets, we generate feature embeddings that facilitate efficient and accurate image retrieval. Our contributions in this work are threefold. First, we integrate EsViT into the histopathology image retrieval pipeline, leveraging self-supervised learning to enhance feature representation without the need for large annotated datasets. Second, we propose a comprehensive retrieval pipeline that includes feature extraction, feature indexing, and similarity-based retrieval. Finally, we employ an optimized k-nearest neighbor (k-NN) search mechanism to efficiently retrieve the most relevant histopathology images based on learned feature embeddings, ensuring accurate and scalable image retrieval.

The paper is structured as follows: Section II describes the proposed methodology, including model architecture, dataset details, and retrieval mechanism. Section III presents experimental results and performance comparisons. Finally, Section IV concludes the paper with future directions.

II. Methods

This section details our proposed histopathology image retrieval system based on Efficient Self-Supervised Vision Transformers (EsViT). The objective of the proposed system is to enable efficient and accurate retrieval of histopathology images based on visual similarity. The workflow consists of four stages. First, feature extraction is performed using an EsViT model, which generates compact feature embeddings from input histopathology images. Second, indexing is conducted by storing extracted feature vectors in a structured database to facilitate fast search and retrieval. Third, query processing involves passing a new histopathology image through the EsViT model to obtain its corresponding feature representation. Finally, retrieval is performed by applying a k-nearest neighbor (k-NN) search to identify the most similar images from the indexed dataset. The retrieval results are ranked based on similarity scores, and the most relevant images are returned to the user. An illustration of the proposed method is shown in Figure 1.

A. Efficient Self-Supervised Vision Transformers

Traditional CNNs process images using hierarchical feature extraction with local receptive fields, limiting their ability to model long-range dependencies. Vision Transformers (ViTs) address this limitation by treating images as sequences of patches, enabling global context modeling. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, ViTs first tokenize the image into non-overlapping patches of size $P \times P$, forming a sequence of $N = \frac{HW}{P^2}$ patches:

$$X = [x_1, x_2, \ldots, x_N], \quad x_i \in \mathbb{R}^{P^2 C}$$

Each patch embedding is obtained by linear projection through a shared learnable matrix W_e:

$$Z_0 = XW_e + E_{\text{pos}}$$

where E_{pos} is the positional encoding to retain spatial information. The embeddings are processed throughmulti-head self-attention (MHSA) layers, where the attention weights are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are the *query, key, and value* matrices obtained by linear transformations of the input, and d_k is the dimensionality of the key vectors.

EsViT extends this standard ViT framework by incorporating hierarchical token merging, where spatially adjacent patches are progressively merged at different layers, reducing the token count while enriching feature representations:

$$Z_{\ell+1} = f_{\text{merge}}(Z_\ell)$$

where f_{merge} represents the hierarchical merging operation, which down-samples tokens while preserving high-level semantics. This approach significantly improves computational efficiency without compromising feature expressiveness.

Additionally, EsViT is pretrained in a self-supervised manner, combining region matching and contrastive learning. Given two augmented views I_a, I_b of the same image, the model is trained to maximize feature similarity between corresponding patches while minimizing similarity between different images:

$$\mathcal{L}_{\text{contrast}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f_a^i, f_b^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f_a^i, f_c^j)/\tau)}$$

where $\text{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$ is the *cosine similarity*, and τ is the temperature scaling factor. This self-supervised pretraining allows EsViT to learn domain-specific representations without requiring labeled data, making it well-suited for histopathology image retrieval.

B. Feature Extraction Using EsViT

Given an input histopathology image I, EsViT generates a *D-dimensional feature vector* $f(I)$ that captures its morphological characteristics. The feature extraction process consists of three steps:

1) **Hierarchical Transformer Encoding:**
   - The image is patch-tokenized and passed through the multi-stage hierarchical ViT backbone.
   - Each layer progressively aggregates spatial information using attention mechanisms, leading to a multi-scale feature representation:

$$Z_\ell = \text{MHSA}(Z_{\ell-1}) + Z_{\ell-1}$$

   where Z_ℓ represents the patch embeddings at layer ℓ.
2) **Global Feature Pooling:**
   - The patch embeddings are aggregated using average pooling to form a global image descriptor:

$$f(I) = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

   This ensures that the final feature vector encapsulates both local and global histopathological structures.
3) **Feature Normalization:**
   - To ensure scale invariance and improve retrieval performance, the extracted feature vector is L2-normalized:

$$\hat{f}(I) = \frac{f(I)}{\|f(I)\|}$$

   This normalization step ensures that feature comparisons are based solely on direction, making cosine similarity an effective retrieval metric.

The final normalized feature vectors serve as compact and discriminative representations, capturing the morphological patterns and spatial relationships necessary for content-based histopathology image retrieval.

C. Dataset Preprocessing and Augmentation

To fine-tune the EsViT model for histopathology image retrieval, we utilize three publicly available datasets: Breast Cancer Histopathology (BracS) [17], Colorectal Cancer Histopathology (CRC) [18], and BreAst Cancer Histology (BATCH) [19]. These datasets contain a diverse set of histopathology images that exhibit variations in staining, magnification, and tissue morphology. To ensure consistency between data sets, all images are resized at a fixed resolution of 224×224 pixels. The pixel values are normalized to the range [0,1] to standardize the input distributions.

Since self-supervised training benefits from diverse transformations, we apply a series of data augmentation techniques, including random cropping and resizing, color jittering, Gaussian blurring, and horizontal flipping. These augmentations help improve the model's ability to generalize across different staining protocols and imaging conditions. For effective training and evaluation, the dataset is split into three subsets: 80% for training, 10% for validation, and 10% for testing. The training set is used to fine-tune EsViT on histopathology images, while the validation set is utilized for hyperparameter tuning. The test set is held out for final performance evaluation to ensure unbiased assessment of the retrieval system.

D. Feature Indexing

After extracting feature vectors for all images in the dataset, the representations are stored in a structured index using FAISS (Facebook AI Similarity Search). FAISS is an optimized similarity search library that allows fast retrieval of high-dimensional feature embeddings. By leveraging FAISS's efficient data structures, the system can handle large-scale histopathology repositories while maintaining low query latency.

E. Query Processing and Similarity Search

Given a query image I_q the retrieval process begins by computing its EsViT feature vector $f(I_q)$. The system then performs a cosine similarity search to measure the closeness of the query image to indexed images. The similarity between the query image and a database image I_i is computed as:

$$Similarity(I_q, I_i) = \frac{f(I_q) \cdot f(I_i)}{\|f(I_q)\|\|f(I_i)\|} \quad (1)$$

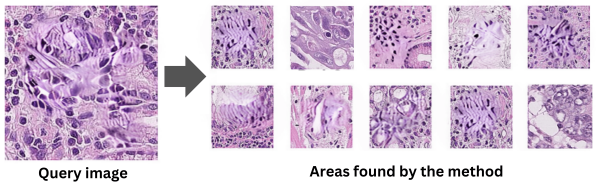

Fig. 2. Illustration of the content-based image retrieval process. The left side shows the query histopathological image, while the right side displays the most similar image regions retrieved by the method.

where I_q and $f(I_q)$ are the feature vectors of the query and database images, respectively. A k-nearest neighbor (k-NN) search is conducted to retrieve the top-k most similar images, which are ranked based on similarity scores. The retrieved images are then displayed to the user, facilitating visual comparison and case-based reasoning. Example of query processing and similarity search in histopathological images is shown in Figure 2.

F. Implementation Details

The system is implemented using PyTorch for deep learning and FAISS [20] for efficient retrieval. The EsViT model is initialized with pretrained weights from ImageNet and fine-tuned on histopathology datasets using the Adam optimizer with a learning rate of 1e-4. The batch size is set to 32, and training is conducted on an NVIDIA A100 GPU, requiring approximately 12 hours for fine-tuning. To ensure stable training, a learning rate warm-up strategy is employed for the first five epochs, followed by a cosine decay schedule.

III. Results and Discussion

To evaluate the retrieval performance of our system, we utilize a set of standard metrics commonly used in content-based image retrieval tasks. Mean Average Precision at k (mAP@k) is employed to quantify the mean of the average precision scores for the top-k retrieved images across all queries, providing an overall measure of ranking accuracy. Additionally, Precision@k is computed to determine the proportion of relevant images within the top-k retrieved results, reflecting the system's ability to prioritize relevant matches. To further assess ranking quality, we use Normalized Discounted Cumulative Gain (nDCG@k), which assigns higher relevance scores to correctly retrieved images that appear earlier in the ranked list, ensuring that more relevant cases are prioritized.

Beyond accuracy, we analyze computational efficiency by measuring feature extraction time, which captures the time required to generate feature representations from input images, and retrieval latency, which quantifies the total time taken to search and rank the most similar images from the indexed database.

A. Quantitative Results

Our proposed EsViT-based retrieval system achieves the highest mAP@10 across all datasets, showing a 9.5% improvement over SimCLR and 13-16% improvement over traditional CNN-based methods. Additionally, the retrieval latency of our system is significantly lower (8.9 ms) compared to SimCLR (16.7 ms) and DenseNet121 (15.3 ms), making it a scalable solution for large-scale histopathology repositories.

One of the key advantages of EsViT is its ability to learn representations from unlabeled data. To quantify the impact of self-supervised pretraining, we compare EsViT with and without pretraining. Table 2 shows that pretraining on ImageNet improves retrieval performance by 5-7% across all metrics, highlighting the benefits of self-supervised representation learning.

IV. Conclusion and Discussion

In this work, we present an efficient histopathology image retrieval system based on EsViT, leveraging self-supervised learning to extract discriminative feature representations without requiring large labeled datasets. First, Vision Transformers (ViTs) process images as sequences of patches, enabling them to model long-range dependencies critical for analyzing intricate tissue structures. Second, self-supervised learning allows EsViT to learn domain-specific representations without requiring large labeled datasets, addressing one of the major challenges in medical image retrieval. Third, EsViT's hierarchical multi-stage design enhances feature expressiveness while maintaining low computational cost.

Experimental results demonstrate that EsViT outperforms CNN-based approaches, achieving up to 16% higher retrieval accuracy in mAP@10, Precision@10, and nDCG@10, while maintaining lower feature extraction time (7.5 ms) and retrieval latency (8.9 ms). The experimental results demonstrate that EsViT significantly outperforms CNN-based models in histopathology image retrieval, achieving superior retrieval accuracy and computational efficiency. This efficiency gain is attributed to optimized token processing, which reduces redundant computations while preserving discriminative features essential for content-based retrieval. Qualitative analysis further confirms that EsViT retrieves morphologically relevant histopathology images, highlighting its potential for clinical decision support and pathology research.

However, several areas for improvement remain. The model currently relies on ImageNet pretraining, which may introduce domain biases due to differences between natural images and histopathology images. Future work should explore pretraining on large-scale histopathology datasets to improve domain adaptation. Additionally, the retrieval system employs cosine similarity, which, while effective, may not fully capture

TABLE I
RETRIEVAL PERFORMANCE COMPARISON ON HISTOPATHOLOGY DATASETS

Model	Dataset	mAP@10 ↑	Precision@10 ↑	nDCG@10 ↑	Feature Extraction Time (ms) ↓	Retrieval Latency (ms) ↓
ResNet50 + k-NN	BracS	62.3	65.1	71.2	9.8	14.2
ResNet50 + k-NN	CRC	58.4	61.3	68.5	9.5	13.8
ResNet50 + k-NN	BATCH	60.2	63.7	70.1	9.6	13.9
DenseNet121 + k-NN	BracS	65.2	68.7	74.5	11.1	15.3
DenseNet121 + k-NN	CRC	60.9	64.4	70.8	10.8	14.9
DenseNet121 + k-NN	BATCH	63.1	67.0	72.9	10.9	15.1
SimCLR + k-NN	BracS	68.9	72.3	77.8	14.2	16.7
SimCLR + k-NN	CRC	63.5	67.2	73.1	13.8	16.4
SimCLR + k-NN	BATCH	65.8	70.1	75.3	14.0	16.5
EsViT + k-NN	BracS	78.4	81.2	85.6	7.5	9.1
EsViT + k-NN	CRC	74.9	77.6	83.2	7.1	8.7
EsViT + k-NN	BATCH	79.1	82.4	86.3	7.2	8.9

TABLE II
EFFECT OF SELF-SUPERVISED PRETRAINING ON RETRIEVAL PERFORMANCE.

Model	Pretraining	mAP@10 ↑	Precision@10 ↑	nDCG@10 ↑
EsViT (No Pretraining)	None	72.1	74.5	80.2
EsViT (Pretrained on ImageNet)	Self-Supervised	**78.4**	**81.2**	**85.6**

complex morphological variations. Contrastive learning-based retrieval strategies or learnable similarity metrics could further refine ranking accuracy. Another promising direction is the development of a multi-modal retrieval framework that integrates histopathology images with clinical metadata, providing more comprehensive search capabilities and improving clinical applicability.

In summary, this work demonstrates the effectiveness of self-supervised vision transformers for histopathology image retrieval, offering an accurate, scalable, and efficient alternative to CNN-based methods. By reducing reliance on labeled data and improving retrieval efficiency, EsViT-based retrieval systems can contribute to advancements in AI-assisted pathology and precision medicine, supporting both clinical workflows and pathology research.

ACKNOWLEDGMENT

This research has been partially supported by the Flanders AI Research Programme grant no. 174K02325 and by the Croatian Science Foundation under project number IP-2024-05-9492. The authors thanks International Medical Center Priora for the specific types of support provided by the hospital.

REFERENCES

[1] W. He, T. Liu, Y. Han, W. Ming, J. Du, Y. Liu, Y. Yang, L. Wang, Z. Jiang, Y. Wang, J. Yuan, and C. Cao, "A review: The detection of cancer cells in histopathology based on machine vision," *Computers in Biology and Medicine*, vol. 146, p. 105636, 2022.

[2] Z. Tabatabaei, F. P. Bueno, A. Colomer, J. O. Moll, R. Molina, and V. Naranjo, "Advancing content-based histopathological image retrieval pre-processing: A comparative analysis of the effects of color normalization techniques," *Applied Sciences*, 2024.

[3] J. de Matos, S. T. M. Ataky, A. de Souza Britto, L. E. S. de Oliveira, and A. L. Koerich, "Machine learning methods for histopathological image analysis: A review," *ArXiv*, vol. abs/2102.03889, 2021.

[4] M. M. Abdelsamea, U. Zidan, Z. Senousy, M. M. Gaber, E. A. Rakha, and M. Ilyas, "A survey on artificial intelligence in histopathology image analysis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, 2022.

[5] M. Hadid, Q. M. Hussein, Z. T. Al-qaysi, M. Ahmed, and M. M. Salih, "An overview of content-based image retrieval methods and techniques," *Iraqi Journal for Computer Science and Mathematics*, 2023.

[6] A. Riasatian, M. Babaie, D. Maleki, S. Kalra, M. Valipour, S. Hemati, M. Zaveri, A. Safarpoor, S. Shafiei, M. Afshari, *et al.*, "Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides," *arXiv preprint arXiv:2101.07903*, 2021.

[7] B. T. Hung and S. Pramanik, "Content-based image retrieval using multi deep neural networks and k-nearest neighbor approaches," *Research Square*, 2023.

[8] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "How deeply to fine-tune a convolutional neural network," *Applied Sciences*, vol. 10, no. 10, p. 3359, 2020.

[9] E. Eslami and H.-B. Yun, "Attention-based multi-scale convolutional neural network (a+mcnn) for multi-class classification in road images," *Sensors*, vol. 21, no. 15, p. 5137, 2021.

[10] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "A survey on graph-based deep learning for computational histopathology," *arXiv preprint arXiv:2107.00272*, 2021.

[11] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Generative adversarial networks in digital pathology and histopathological image processing: A review," *Medical Image Analysis*, vol. 67, p. 101813, 2021.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *Journal of Digital Imaging*, vol. 36, no. 1, pp. 1–18, 2023.

[15] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, "Efficient self-supervised vision transformers for representation learning," *ArXiv*, vol. abs/2106.09785, 2021.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.

[17] I. di Calcolo e Reti ad Alte Prestazioni, "Breast cancer histopathology images."

[18] Warwick, "Colorectal cancer histopathology images."

[19] T. e. C. I. T. Universidade do Porto, Instituto de Engenharia de Sistemas e Computadores, I. de Investigação, and P. Inovação em Saúde (i3S), "Breast cancer histology images."

[20] Meta, "Faiss."