FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY OSIJEK
JOSIP JURAJ STROSSMAYER UNIVERSITY OF OSIJEK
MEDITRAIN
HrZZ Croatian Science Foundation
FACULTY OF ENGINEERING AND ARCHITECTURE
GHENT UNIVERSITY
medi-train.github.io

Efficient Self-Supervised Vision Transformers for Histopathology Image Retrieval

Marija Habijan, Petar Nakić, Irena Galić, Danijel Marinčić, Josip Samardžić, Aleksandra Pižurica

✉ marija.habijan@ferit.hr

1 Motivation and Problem Statement

Histopathology is essential for cancer diagnosis and prognosis, where microscopic tissue images support the detection of abnormalities. With the increasing availability of digitized whole-slide images, efficient image retrieval systems are needed to aid clinical decision-making, and research. With the increasing availability of digitized whole-slide images, efficient image retrieval systems are needed to aid case-based reasoning, clinical decision-making, and research. Nevertheless, histopathology image retrieval is limited by two main factors:

1. Existing CNN-based methods rely heavily on large annotated datasets, which are scarce and expensive to obtain in medical imaging.
2. Models struggle with scalability and often fail to capture long-range tissue dependencies, leading to suboptimal retrieval accuracy and efficiency.

This motivates the development of retrieval systems that can learn from unlabeled data, handle large repositories, and deliver fast, accurate query responses for clinical use.

2 Proposed Method

We propose a retrieval system based on Efficient Self-Supervised Vision Transformers (EsViT). EsViT extends Vision Transformers with hierarchical token merging and self-supervised pretraining, enabling compact yet expressive image representations. The system pipeline consists of:

1. Feature extraction – EsViT encodes histopathology images into discriminative feature vectors.
2. Feature indexing – Extracted features are stored using the FAISS similarity search library.
3. Query processing – A query image is passed through EsViT to obtain its feature embedding.
4. Similarity retrieval – k-nearest neighbor search with cosine similarity identifies the most relevant cases.

This system ensures efficient and scalable retrieval with reduced reliance on labeled datasets and an illustration of the proposed method is shown in Figure 1.

3 Results and Discussion

Across all datasets, EsViT consistently outperformed CNN-based methods (ResNet50, DenseNet121) and contrastive learning baselines (SimCLR). It achieved mAP@10 of 78.4% (BracS), 74.9% (CRC), and 79.1% (BATCH), with Precision@10 exceeding 81%, representing up to 16% improvement over CNNs (Table 2). Ablation studies further demonstrated the importance of self-supervised pretraining. Without pretraining, EsViT achieved 72.1% mAP@10, while ImageNet pretraining improved performance by 5–7% across all metrics (Table 1). Qualitative analysis revealed that the system retrieves morphologically relevant regions that align with clinical expectations (Figure 2).

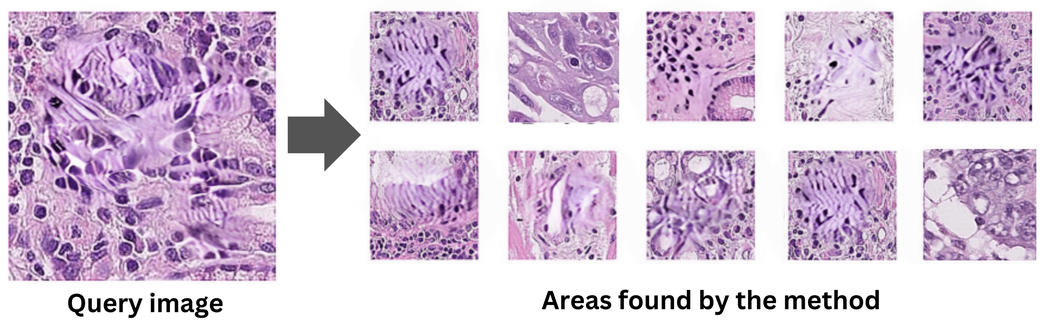

Figure 1: Illustration of the content-based image retrieval process.

Table 1: 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**

4 Conclusions

This work demonstrates that self-supervised vision transformers provide an accurate, scalable, and efficient solution for histopathology image retrieval. By reducing dependence on large labeled datasets and achieving superior retrieval accuracy, EsViT offers strong potential for integration into clinical workflows and pathology research.

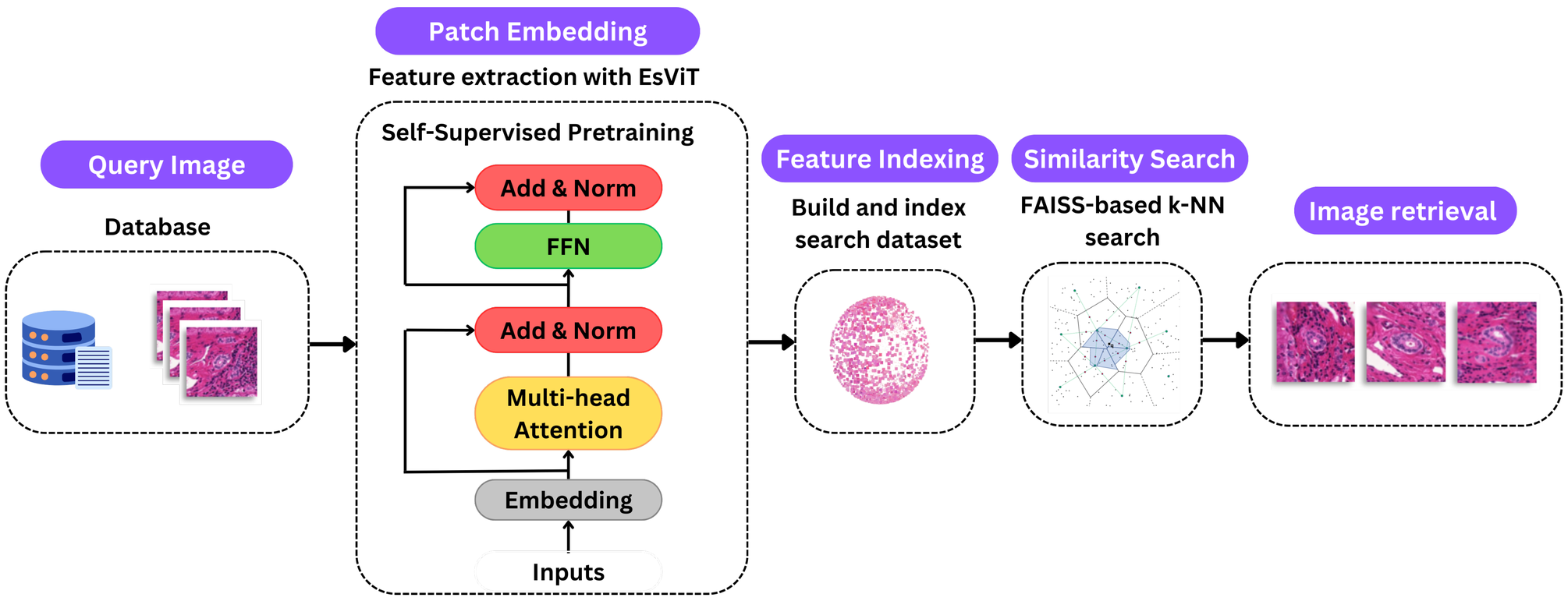

Figure 2: Overview of the proposed EsViT-based histopathology image retrieval system.

Table 2: 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

Published in: M. Habijan, P. Nakić, I. Galić, D. Marinčić, J. Samardžić, and A. Pizurica. *Efficient self-supervised vision transformers for histopathology image retrieval. EUSIPCO 2025.*

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.