Machine Learning in Applications
AM03 Project Report

Barbieri Fabio, Di Giorgio Vittorio, Ferrigno Antonio, Scoleri Maria Rosa
Politecnico di Torino

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Machine Learning in Applications AM03 Project Report

***Abstract*—This study evaluates the efficacy of different self-supervised learning (SSL) methods for feature extraction from a Colorectal Cancer Whole Slide Images dataset (CRC-WSIs). We compare these SSL approaches against conventional transfer learning techniques using ImageNet pre-trained Convolutional Neural Networks (CNNs). To ensure a fair and straightforward comparison, we assess each model's representational capabilities by using their learned embeddings as input to a linear SVM for a cancer detection task. Our investigation involves a representative set of SSL methods, including state-of-the-art architectures, training strategies, and data augmentation techniques. The selected methods include a Variational Autoencoder, a rotation agnostic framework that includes Vision Transformers (PathDino), and a contrastive learning method (Bootstrap Your Own Latent, BYOL). We find that PathDino, especially when fine-tuned, outperforms the other models in extracting meaningful features. Interesting results are given by BYOL as well, prompting an interesting discussion on the augmentation techniques for medical images.**

For more information and code, please visit our GitHub repository.

I. Introduction

The digital pathology field has greatly benefited from the advancements in Deep Learning in recent years, particularly in the application of deep learning models for analyzing Whole Slide Images (WSIs). These high-resolution images provide a high amount of information for diagnosing and studying diseases such as colorectal cancer (CRC). However, effective use of deep learning models in this domain faces several challenges, including the scarcity of large-scale annotated datasets and the high computational resources required.

Traditional approaches have relied heavily on transfer learning, where models pre-trained on large natural image datasets like ImageNet [2] are fine-tuned for specific pathology tasks. While this method has shown promise, it raises questions about the appropriateness of using features learned from natural images for medical imaging tasks, given the significant domain differences. Self-supervised learning (SSL) has emerged as a powerful alternative, offering the potential to learn meaningful representations from unlabeled medical imaging data. SSL methods can leverage the vast amounts of unlabeled WSIs available in pathology archives, potentially capturing domain-specific features more relevant to pathology tasks than those learned from natural images [3]. However, the applicability of typical SSL augmentation techniques to WSIs is problematic. Many SSL methods rely on augmentations such as rotations, flips, blurs, and color jittering, which are effective for natural images but may not be appropriate or beneficial for histopathological images. For instance, blurring can make it difficult to clearly see the size, shape, and chromatin pattern of tumor cell nuclei. The orientation of tissue structures may carry diagnostic significance, so arbitrary rotations could potentially distort important features. Similarly, color variations in histopathology images are often meaningful and tied to staining processes, making indiscriminate color augmentations potentially counterproductive. These concerns highlight the need for a careful evaluation of SSL methods and their augmentation strategies in the context of WSIs.

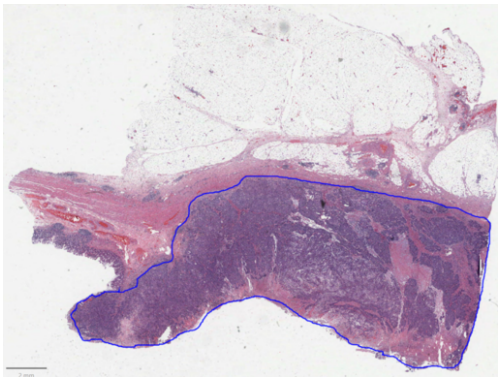
The purpose of this study is to conduct a comprehensive evaluation of various SSL methods for feature extraction from Colorectal Cancer Whole Slide Images (CRC-WSIs), comparing their efficacy against conventional transfer learning approaches. By focusing on a diverse set of SSL techniques, including Variational Autoencoders (VAEs [4]), rotation-agnostic frameworks with Vision Transformers (PathDino [5]), and contrastive learning methods like Bootstrap Your Own Latent (BYOL [1]), we aim to provide insights into the most effective approaches for learning representations from WSIs. Our investigation is designed to address several key questions: How do different SSL methods compare in their ability to extract meaningful features from CRC-WSIs? How do the relative performances relate to model training strategies and architectures? Can SSL approaches outperform traditional transfer learning techniques using ImageNet pre-trained Convolutional Neural Networks (CNNs) in the context of digital pathology?

To ensure a fair and interpretable comparison, we evaluate these methods by using their learned embeddings as input to a linear Support Vector Machine (SVM) for a cancer detection task. This approach allows us to assess the quality of the learned representations directly, without employing complex downstream architectures.

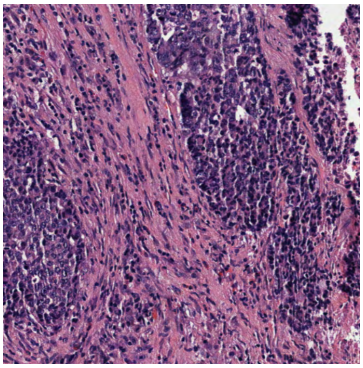
II. Background

Representation learning has become a cornerstone in medical imaging, especially in relation to its application to digital pathology. Over the years, several methods have been developed and tested in this field, including techniques of unsupervised learning and self-supervised learning.

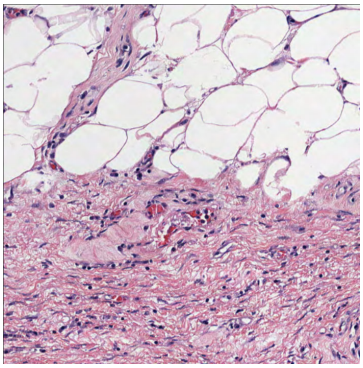
Variational Autoencoders (VAEs) have emerged as a powerful tool in the field of representation learning for medical imaging. For example, the work presented in [6] exploits the encoding of VAEs to analyze histopathological images resulting from breast tissue biopsies to classify the tumor as cancerous or benign. The research presented in [7] introduces a VAE-based approach to compress and decompress cancer pathology slides verifying that the compression still allows to maintain accuracy in clinical validation tasks. The same conclusion can be found in [8]. VAEs, however, are not without limitations: in some scenarios, they struggle to represent fine



(a) Original WSI image



(b) Cancerous patch



(c) Non-cancerous patch

Fig. 1: Figure (a) shows an example of a WSI image taken from the CRC-WSIs dataset. The blue border delimits the cancerous region of the image. (b) represents a cancerous patch taken from the inside of the blue border of the original image, and (c) shows a non-cancerous patch.

details or sharp transitions in the data, this is why more advanced methods have also been developed.

A recent advancement in the representation learning of medical images is the use of models based on Vision Transformer (ViT) [9]. Thanks to the attention mechanism [10], transformers have proven to be capable of learning long-range dependencies and spatial correlations both in Natural Language Processing and in Computer Vision. The review presented in [11] shows how ViTs are used in different medical image analysis tasks, including classification, segmentation, detection, and clinical report generation. ViTs however, can fail to account for the unique characteristics of histopathological images. They also require large amounts of data to perform optimally and can face issues of overfitting. To address these challenges, the authors of [5] present a model called **PathDino**, which is a new ViT-based approach that is designed specifically for histopathological images. PathDino is a lightweight and compact feature extractor built with five transformer blocks. Accompanied by a novel patch selection method and a rotation augmentation technique, PathDino demonstrates superior performance and effectively manages to reduce overfitting. This work is particularly relevant to our study and will be further analyzed and tested in relation to our dataset.

To further enhance the quality of learned representations in medical images, contrastive learning models have also been explored. The work presented in [12], for example, proposes different strategies for extending the contrastive learning framework for the segmentation of volumetric medical images with limited annotation. The authors of [13] highlight the need to develop a method capable of learning effective representations when labeled data are scarce, and they introduce novel contrasting strategies that leverage structural similarity across volumetric medical images.

III. Materials and Methods

In this section, we provide a comprehensive description of the materials and methodologies used in our investigation. Starting with the dataset description and the extracted patches, moving forward to the baselines, and finally to the self-supervised learning methods we tested.

A. Dataset

Our study focuses on a set of Colorectal Cancer Whole Slide Images (CRC-WSIs) from 24 patients. Each patient's data consists of an SVS image and a corresponding XML file containing Region-of-Interest (ROI) coordinates which delineate areas of cancer presence. To create a suitable dataset, we processed the images to extract 512x512 pixel patches at magnification level 1 - which has a 0.252 downsampling factor. The extracted patches are then further processed to obtain a dataset of (patch, label) pairs where labels are boolean values encoding cancer presence in the relative patch. We used the Ray Casting algorithm as a labeling method in order to determine if patch vertices fall within the cancerous region. An example of a WSI image and patches taken from our dataset is shown in Figure 1. The obtained dataset has the following characteristics:

- Total samples: 13878
- Cancerous samples: 6880
- Non-cancerous samples: 6998

As we can see the obtained dataset is fairly balanced, but it is relatively small compared to typical WSI settings ([14]), emphasizing the need for suitable augmentation techniques.

B. Baselines

The baselines chosen to test the strength of SSL methods are the DenseNet121 [15] and ResNet50 [16] networks. DenseNet-based models have been used extensively in the field of medical imaging for tasks such as classification [17], [18], pattern recognition, image segmentation, and object detection [19]. We selected DenseNet121 as a baseline model due to its strong representational capabilities while being less computationally expensive compared to deeper architectures. For this study, we use the DenseNet121 model pre-trained on ImageNet, and no specific hyperparameter tuning was performed.

ResNet50 [16] is another widely used CNN architecture that is deeper than DenseNet121, consisting of 50 layers. It

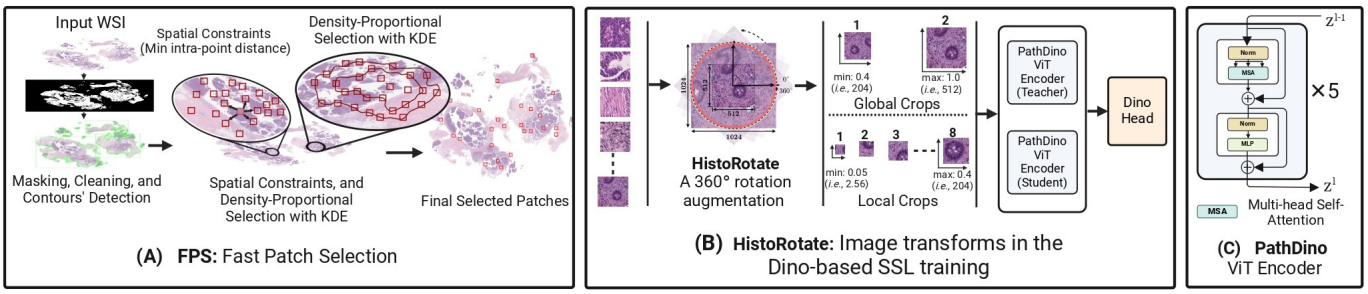


Fig. 2: The figure shows the WSI Analysis Pipeline and it is taken from [5]. (A) The Fast Patch Selection (FPS) method chooses a group of representative patches while maintaining their spatial arrangement. (B) HistoRotate is a 360°rotation augmentation for histopathology model training that helps the representations of the models become rotation invariant. (C) PathDino is a compact histopathology Transformer with five small vision transformer blocks and about 9 million parameters

is built on the concept of residual learning, where shortcut connections, or "skip connections," allow for the construction of very deep networks by alleviating the vanishing gradient problem. In the context of digital pathology, ResNet50's deep architecture and efficient learning mechanism make it particularly well-suited for capturing the intricate patterns and multi-scale features present in Whole Slide Images. Its proven track record in transfer learning scenarios, where models pre-trained on large datasets like ImageNet are adapted to specific medical imaging tasks, further justifies its selection as a strong baseline for our comparative study [20].

C. BYOL

Bootstrap Your Own Latent (BYOL) is a self-supervised learning framework that learns image representations without the need for negative pairs, differing from traditional contrastive learning methods. BYOL employs two neural networks, an online network and a target network, which learn collaboratively via a bootstrapping mechanism.

The online network comprises an encoder, a projector, and a predictor, which process input images to produce embeddings that are aligned with the outputs of the target network. The target network, which does not include a predictor, shares a similar architecture but updates its parameters as a moving average of the online network's parameters, allowing for gradual evolution.

The core of BYOL's learning mechanism is its loss function, which aims to align the outputs of the predictor in the online network with the target network. Given two augmented views of an image, v and v', the loss function is defined as:

$$\mathcal{L}_{BYOL} = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2}.$$

In this loss function:

- $q_\theta(z_\theta)$ is the output of the predictor network in the online network.
- z'_ξ is the embedding generated by the target network from an augmented view v' of the input image.
- $\langle q_\theta(z_\theta), z'_\xi \rangle$ denotes the dot product between the predictor's output and the target network's embedding.
- $\|q_\theta(z_\theta)\|_2$ and $\|z'_\xi\|_2$ are the L2 norms of the vectors $q_\theta(z_\theta)$ and z'_ξ, respectively.

This loss measures the cosine similarity between the normalized outputs of the online network's predictor and the target network, promoting similarity between representations of different augmented views of the same image.

To avoid representation collapse, where the model produces identical outputs for all inputs, BYOL uses an asymmetric design. The target network serves as a stable reference with slowly evolving parameters, while the predictor in the online network drives the learning of diverse and meaningful features, preventing convergence to trivial solutions.

D. PathDino

PathDino is a self-supervised learning model specifically designed to enhance image representation learning in digital pathology. A schematic representation of the model's architecture and process can be found in Figure 2. It employs a compact transformer-based architecture with approximately 9 million parameters, making it efficient for deployment in systems with limited computational resources. PathDino's framework includes an architecture with five small vision transformer blocks, a Fast Patch Selection (FPS) method, and a rotation augmentation technique called HistoRotate.

Key Components are:

- **Fast Patch Selection (FPS):** Selects representative patches from whole slide images (WSIs) while preserving spatial distribution. The process consists of creating a smaller thumbnail image from the WSI, as well as a tissue mark to identify areas of interest. Potential patch locations are determined based on contours in the tissue mark. Then, Kernel Density Estimation (KDE) is used to calculate the density of these potential patches and select them. A minimum distance is maintained between selected patches to avoid choosing too many patches from densely packed regions. Finally, the selected patches are mapped back to their original positions in the high-resolution WSI.
- **HistoRotate:** A 360-degree rotation augmentation method that enhances learning without altering the contextual information of histopathology images. Two rotations are considered: an angle is sampled from a continuous uniform distribution in the range $[0, 360]$

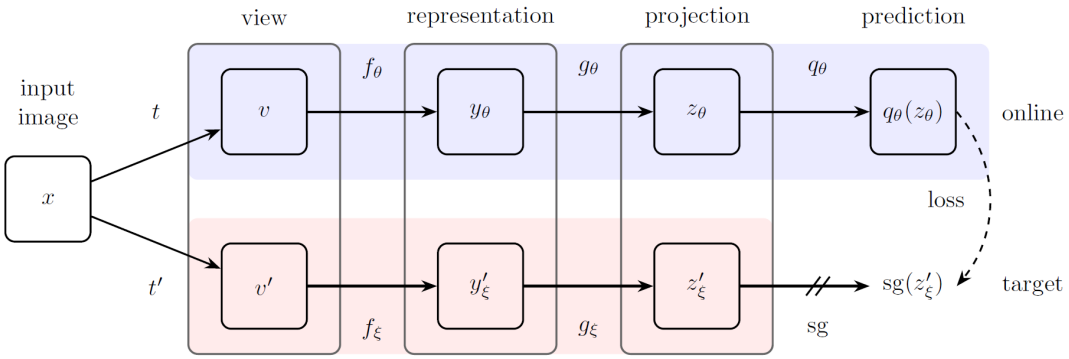


Fig. 3: BYOL Architecture from [1]. BYOL minimizes a similarity loss between the prediction part of the online network $q_\theta(z_\theta)$ and the output of the target network after stop gradient $sg(z'_\xi)$. θ symbolizes the trained weights and ξ are an exponential moving average of θ. After training, everything except f_θ is discarded and y_θ is used as the image representation.

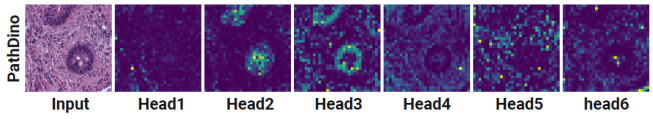


Fig. 4: Visualization of attention maps generated by different attention heads in the PathDino model [21]. The input histopathology image is shown on the left, followed by the attention maps from each of the six attention heads. These maps illustrate how each head in the transformer model focuses on different elements of the tissue structure, highlighting the model's ability to capture critical features across multiple levels of abstraction in the input image.

degrees. A second discrete rotation is selected from the set $\Theta = \{90, 180, 270, 360\}$.

- **Vision Transformer Blocks:** The model architecture is composed of five compact Vision Transformers blocks. Each block is made with a multi-head self-attention (MSA) layer, Layer Norm (LN), and a Multilayer Perceptron (MLP):

$$\mathbf{z}_i^l = MLP(LN(MSA(\mathbf{z}_i^{l-1}))) + \mathbf{z}_i^{l-1} \tag{1}$$

where $l = 1, ..., L$ and $n = 1, ..., N$ with $L = 5$ blocks and N represents the total input patches.

The attention mechanism in PathDino plays a crucial role in capturing critical features from histopathology images by focusing on different regions of the input image through multiple attention heads. Each of the five attention heads in PathDino's vision transformer model generates an attention map that highlights distinct tissue structures and patterns. Figure 4 illustrates the attention maps generated by each head.

E. ResVAE

Variational Autoencoders [4] have proved to be one of the most effective tools for representation learning. For the purpose of our work we employ a variational autoencoder based on a ResNet [16] architecture (ResVAE) [1]. In particular, the encoder is a modified version of a ResNet18 in which the series of residual blocks with skip connections performs the downscaling of the image. The decoder is a mirror of the encoder and performs the upscaling operation. The VAE employs the reparameterization trick to sample from the latent space during training, utilizes a Mean-Squared Error (MSE) as reconstruction loss, and the KL Divergence to regularize the latent space. The MSE and the KL divergence can be formalized as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{2}$$

$$\text{KL}(q(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{d}(1 + log(\sigma_j^2) - \mu_j^2 - \sigma_j^2) \tag{3}$$

where

- $q(z|x)$ represents the learned latent distribution
- $p(z)$ is the target distribution, usually a standard normal distribution
- $1 + log(\sigma_j^2)$ is the term that penalizes the model if the variance σ_j^2 deviates from 1
- μ_j^2 penalizes the model for having a mean μ that deviates from 0
- σ_j^2 encourages the variance to be close to 1

VAEs require a careful balance between these two losses in order to achieve better performance: minimizing more the reconstruction loss may lead to overfitting but minimizing more the KL divergence could mean that the model does not learn effectively enough the distribution of the training data

F. SVM training

The embeddings extracted from all the aforementioned models are then fed to a linear Support Vector Machine (SVM), which is trained to classify each image as cancerous or non-cancerous. The performance of the SVM reflects the ability of the chosen models to correctly represent the input images and retain the information useful to identify cancerogenous tissue. The SVM was chosen for its simplicity and strong performance in high-dimensional spaces. It is important to underline that all the models presented are trained in a self-supervised manner and used as features extractors. The SVM is the only model that has access to the labeled

[1]The implementation of the variational autoencoder is strongly inspired by the work in this folder.

Model	Model Performance		
	Accuracy	Precision	Recall
Densenet121 (baseline)	0.78	0.78	0.78
ResNet50 (baseline)	0.86	0.87	0.86
ResNet50 (last layer fine-tune)	0.85	0.85	0.85
PathDino (last layer fine-tune)	**0.93**	**0.93**	**0.93**
PathDino (complete fine-tune)	0.84	0.85	0.85
PathDino	0.89	0.89	0.90
BYOL	0.88	0.88	0.88
ResVAE	0.75	0.74	0.74

TABLE I: Models accuracy and macro average precision and recall.

images. t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to the PCA-reduced features for visualization. A scatter plot was generated to visualize the clustering of predicted labels in a 2D space, providing insight into the classifier's performance.

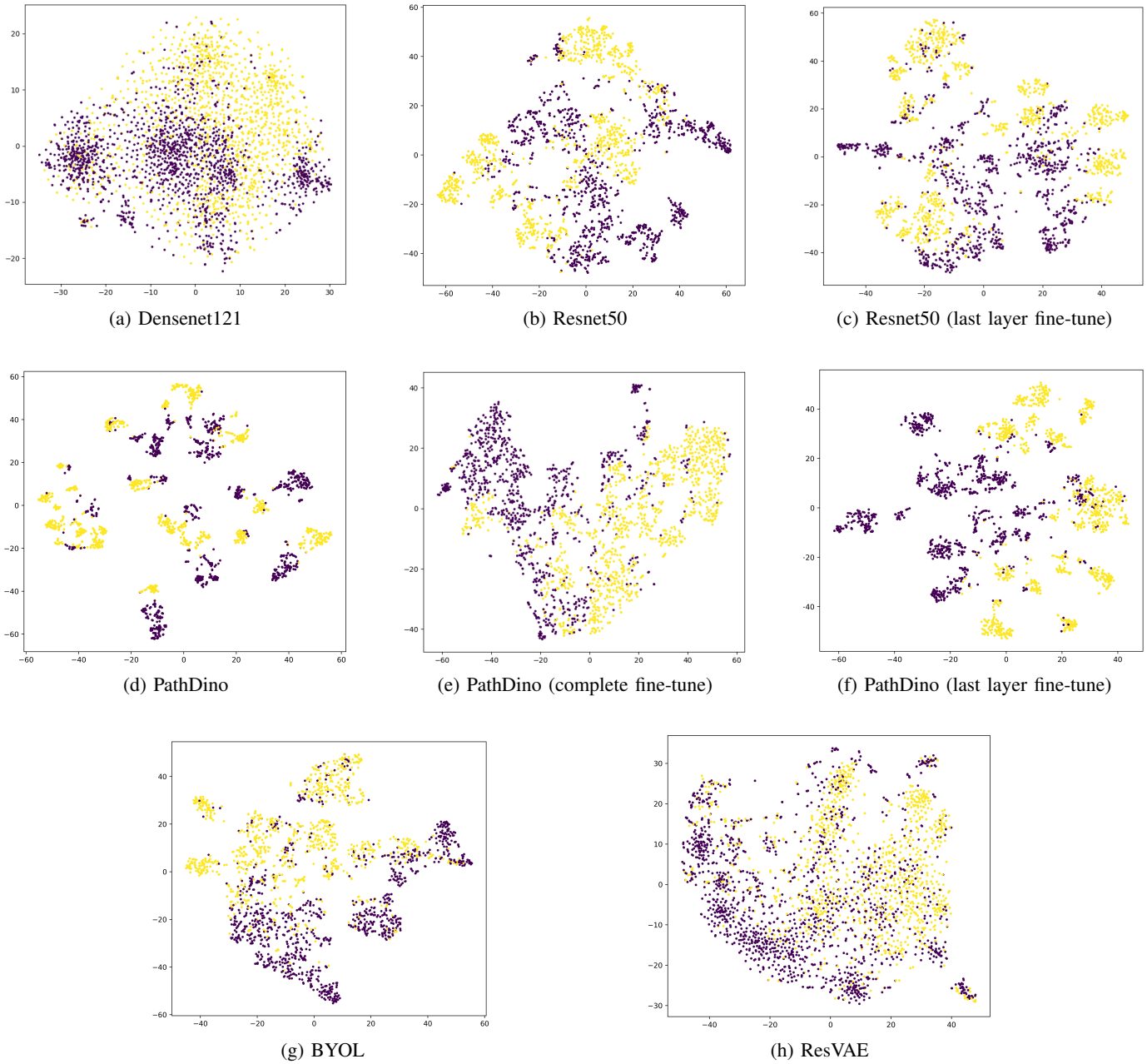
IV. Results and Discussion

In this section, we present and analyze the performance of tested models, as well as the parameters and configurations used to train and evaluate them,

A. Parameters and experiments

- **Baselines**: To benchmark the SSL methods fairly, we employed two main strategies. The first is to compare pre-trained models with pre-trained baselines, the second consists in fine-tuning both models and baselines with different tuning depths. This way we provide a comparative understanding of how well models pre-trained on natural images perform on histopathological data, both with and without task-specific fine-tuning. For DenseNet121, we used no tuning strategy, evaluating only the features extracted from the pre-trained model. For Resnet50, we tested features extracted with the pre-trained model as well as with the fine-tuned model. For the fine-tuning setting, we performed both tuning through the entire network and only on the last fully connected layer. All the tuning experiments were conducted for 10 epochs with the same pre-training optimizer and a learning rate smaller than the one used for pre-training. Finally, we used a cosine annealing scheduling with a weight decay of 0.01.
- **BYOL**: For the BYOL model, we utilize a batch size of 32, 30 training epochs, a learning rate of 3^{-5}, and an image size of 256 pixels. The model is implemented using a ResNet architecture pre-trained on ImageNet, and the optimizer employed for training is Adam.
  The data augmentations applied during training are crucial for learning effective representations. We used transformations inspired by the SimCLR [22] framework, which has proven effective in self-supervised learning settings. The following transformations were applied:
  - A color jitter transformation with parameters set to 0.8 for brightness, contrast, and saturation, and 0.2 for hue, using a strength factor of $s = 1$. This augmentation can encourage the model to learn color-invariant features.
  - A random resized crop to randomly scale and crop the image to the desired size of 256 pixels, with a scale range of (0.2, 1.0). This transformation helps the model learn scale-invariant features by presenting objects at different sizes and positions.
  - A random horizontal flip to flip the image horizontally with a 50% probability, in order to help the model to be invariant to orientation.
  - A random Gaussian blur with a kernel size proportional to the image size (computed as $\frac{\text{IMAGE\_SIZE}}{20*2}+1$) and a standard deviation in the range of (0.1, 2.0), applied with a probability of 10%.
  - A random grayscale transformation applied with a probability of 20%. This forces the model to learn features that are not reliant on color information.
  - A normalization step using mean values of [0.485, 0.456, 0.406] and standard deviations of [0.229, 0.224, 0.225] to match the statistics of the pre-trained ResNet model.

  We perform two experiments: the first one using all the aforementioned augmentations as they were presented, and the second one where we better adapt them to our WSI scenario. We remove the Gaussian blur augmentation because, while it can help the network to focus on more relevant objects and shapes in natural images, in the context of medical patches, it risks erasing potentially significant details. Finally, in the second experiment, we replace the mean and normalization values of the normalization steps with those computed from our datasets. The second experiment, while it does not drastically change the results, allows us to improve the performance of BYOL by 2% with respect to its standard version. The results in Table I already show the improved BYOL performance with the updated augmentations. These transformations, implemented using Kornia's augmentation library, are applied twice to generate two different augmented views of the same image, denoted as `transform1` and `transform2`.
- **PathDino**: the original PathDino-512 from [5] is trained using around 6 million patches from The Cancer Genome Atlas (TCGA, [23]), a batch size of 192, AdamW optimizer and a learning rate of 5^{-4} for 27 epochs. For our work, we perform two fine-tuning experiments: one modifying the weights of the entire network, and one freezing everything but the last ViT block. Our fine-tuning is performed for 10 epochs with the same optimizer and a learning rate smaller than the one used for its pre-training (0.0001 instead of 0.0005). Finally, we used a cosine annealing scheduling with a weight decay of 0.01. Note that this setting fairly compares to fine-tuning strategies applied for the Resnet50 baseline.
- **ResVAE**: our variational autoencoder was trained for 30 epochs with a learning rate of 1^{-5}, Adam optimizer, a *beta* parameter to weight the KLD loss of 1^{-5} and it supports a latent dimension of 128 with an input image



(a) Densenet121 (b) Resnet50 (c) Resnet50 (last layer fine-tune)

(d) PathDino (e) PathDino (complete fine-tune) (f) PathDino (last layer fine-tune)

(g) BYOL (h) ResVAE

Fig. 5: t-SNE visualizations for the different methods. These plots show the structure and separation of the latent space learned by our extraction models. Coherently with the numerical evaluation, the t-SNEs reported here show that the original PathDino and its fine-tuned version (last layer), are the most efficient in separating the representations of cancerous and non-cancerous images.

of 256x256 pixels.

B. Results

The main results of our model are shown in Table I. We choose to evaluate our model using the accuracy and the macro average of precision and recall. To further evaluate the learning efficiency of our models we present their t-SNE, which is used to visually assess how well the models have learned to distinguish between classes by showing the clustering and separation of data points in a two-dimensional space. The t-SNE plots are shown in Figure 5.

C. Discussion

Our study provides several insights into the efficacy of different methodologies for extracting meaningful representations from colorectal cancer whole slide images (WSIs).

Firstly, the findings substantiate the hypothesis that in this domain, self-supervised learning (SSL) methods surpass traditional transfer learning approaches using convolutional neural networks (CNNs) pre-trained on natural images. This is evidenced by the superior performance of SSL models like PathDino and BYOL compared to pre-trained DenseNet121 and ResNet50. The results indicate that features learned directly from histopathological images are more relevant and ef-

fective for cancer detection tasks than those learned from natural images, underscoring the importance of domain-specific feature extraction.

Interestingly, the Variational Autoencoder (VAE) did not follow this trend, exhibiting a lower accuracy (0.75) compared to DenseNet121 (0.78). This underperformance may be attributed to the relatively small size of our dataset and the limited computational resources available. VAEs typically require larger datasets to effectively learn representations, highlighting the challenges of applying such models in resource-constrained settings that are common in medical imaging.

Among the SSL methods evaluated, PathDino, pre-trained on a large dataset of WSIs patches, significantly outperformed both the ImageNet baselines and other SSL methods. This result underscores the critical role of domain-specific pretraining in medical imaging. PathDino's superior performance, achieving an accuracy of 0.89 without fine-tuning and 0.93 with fine-tuning the last ViT block, compared to BYOL's accuracy of 0.88, suggests that its architectural design and pretraining strategy are particularly well-suited for capturing relevant features in histopathological images.

BYOL also demonstrated robust performance. This success is attributable to BYOL's self-supervised learning framework, which allows the model to learn pertinent features directly from the data without relying on negative pairs or explicit labels. The flexibility and robustness of BYOL's approach enable it to effectively capture diverse features from WSIs, even without the domain-specific pretraining that benefits PathDino. The t-SNE visualizations further elucidate these findings; BYOL's representations exhibit clear clustering, effectively distinguishing between cancerous and non-cancerous features. This capability to discern meaningful representations, even in the absence of domain-specific customization, highlights BYOL's potential to generalize across different data domains.

Furthermore, our experiments with BYOL reveal the impact of specific data augmentations on performance. We found that removing Gaussian Blur from the SimCLR and correctly normalizing the images, leads to improved results, with accuracy increasing from 86% to 88%. This indicates that certain augmentations, while generally beneficial for SSL, may not always be optimal for all types of histopathological data, and careful selection of augmentations can significantly influence model performance.

The fine-tuning experiments revealed notable patterns. Fine-tuning the entire network was generally detrimental for both the baselines and most SSL methods, likely due to overfitting on the relatively small dataset. However, fine-tuning only the last layers resulted in significant improvements for PathDino, enhancing accuracy from 0.89 to 0.93, while the Resnet50 baseline did not improve in this scenario. This suggests that the features learned by PathDino are highly transferable and require only minor adjustments to adapt to specific tasks, demonstrating the effectiveness of targeted fine-tuning in leveraging pre-trained models in clinical scenarios.

The exceptional performance of PathDino can be attributed to several factors. The Fast Patch Selection (FPS) method employed for its training likely facilitates efficient and representative sampling of the WSIs, which is crucial when dealing with large images. Furthermore, its pretraining on a substantial WSI dataset enables the model to capture domain-specific features that are highly relevant to the task, enhancing its ability to perform accurately in medical imaging applications.

Lastly, our results partly challenge the prevailing notion in the literature that rotation augmentations are ineffective for self-supervised learning on histological images. The rotation augmentation used in PathDino (HistoRotate) shows significant benefits because it allows the network to learn rotation-invariant representations. This finding further proves that well-designed augmentations tailored to the specific characteristics of histopathological images can enhance the learning of robust representations.

V. Conclusions and Future Works

We conducted a comprehensive evaluation of various self-supervised learning (SSL) techniques for feature extraction from Colorectal Cancer Whole Slide Images (CRC-WSIs). The models under consideration, including BYOL, PathDino, and ResVAE, were compared against traditional transfer learning approaches using pre-trained DenseNet121 and ResNet50 architectures. Our findings demonstrate that the PathDino model, particularly when fine-tuned, outperforms other models, indicating its robustness in capturing meaningful features from histopathological images.

Future work should explore the integration of larger and more diverse datasets to further validate the generalizability of the models tested in this study. It would also be interesting to exploit the Fast Patch Selection and the HistoRotate techniques from the work in [5] in combination with other models and approaches like the contrastive learning that distinguished BYOL. Additionally, domain-specific augmentations, both in the BYOL domain and in other contexts, could enhance performance.

Beyond the current evaluation framework, exploring new downstream tasks, such as the segmentation of cancerous areas or the identification of specific tissue types, could provide additional insights into the utility of these models in clinical settings.

References

[1] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020. [Online]. Available: https://arxiv.org/abs/2006.07733

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[3] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *CoRR*, 2023.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[5] S. Alfasly, A. Shafique, P. Nejat, J. Khan, A. Alsaafin, G. Alabtah, and H. R. Tizhoosh, "Rotation-agnostic image representation learning for digital pathology," 2024. [Online]. Available: https://arxiv.org/abs/2311.08359

[6] H. Guleria, A. Luqmani, H. Kothari, P. Phukan, S. Patil, P. Pareek, K. Kotecha, A. Abraham, and L. Gabralla, "Enhancing the breast histopathology image analysis for cancer detection using variational autoencoder," *International Journal of Environmental Research and Public Health*, vol. 20, p. 4244, 02 2023.

[7] M. S. Nasr, A. Hajighasemi, P. Koomey, P. B. Malidarreh, M. Robben, J. R. Saurav, H. H. Shang, M. Huber, and J. M. Luber, "Clinically relevant latent space embedding of cancer histopathology slides through variational autoencoder based image compression," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2023. [Online]. Available: http://dx.doi.org/10.1109/ISBI53787.2023.10230343

[8] J. Keighley, M. de Kamps, A. Wright, and D. Treanor, "Digital pathology whole slide image compression with vector quantized variational autoencoders," 04 2023, p. 50.

[9] 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," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[11] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," *Medical Image Analysis*, vol. 91, p. 103000, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523002608

[12] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," 2020. [Online]. Available: https://arxiv.org/abs/2006.10511

[13] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," 2021. [Online]. Available: https://arxiv.org/abs/2011.13971

[14] B. Ehteshami Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. M. van der Laak, C. Consortium, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[17] T. Chauhan, H. Palivela, and S. Tiwari, "Optimization and fine-tuning of densenet model for classification of covid-19 cases in medical imaging," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100020, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096821000136

[18] N. Hasan, Y. Bao, A. Shawon, and Y. Huang, "Densenet convolutional neural networks application for predicting covid-19 using ct image," *SN Computer Science*, vol. 2, 09 2021.

[19] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense convolutional network and its application in medical image analysis," *BioMed Research International*, vol. 2022, pp. 1–22, 04 2022.

[20] H. Kim, A. Cosa-Linan, and N. e. a. Santhanam, "Transfer learning for medical image classification: a literature review," *BMC Med Imaging*, 2022.

[21] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[23] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.