Mod Pathol 38 (2025) 100636

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Evaluation of a Task-Specific Self-Supervised Learning Framework in Digital Pathology Relative to Transfer Learning Approaches and Existing Foundation Models

Tawsifur Rahman[a,*], Alexander S. Baras[b], Rama Chellappa[a,c]

[a] Department of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore Maryland; [b] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore Maryland; [c] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland

ARTICLE INFO

Article history:
Received 9 April 2024
Revised 6 September 2024
Accepted 15 October 2024
Available online 23 October 2024

Keywords:
computational pathology
self-supervised learning
spatial-channel attention
weakly supervised learning

ABSTRACT

An integral stage in typical digital pathology workflows involves deriving specific features from tiles extracted from a tessellated whole-slide image. Notably, various computer vision neural network architectures, particularly the ImageNet pretrained, have been extensively used in this domain. This study critically analyzes multiple strategies for encoding tiles to understand the extent of transfer learning and identify the most effective approach. The study categorizes neural network performance into 3 weight initialization methods: random, ImageNet-based, and self-supervised learning. Additionally, we propose a framework based on task-specific self-supervised learning, which introduces a shallow feature extraction method, employing a spatial-channel attention block to glean distinctive features optimized for histopathology intricacies. Across 2 different downstream classification tasks (patch classification and weakly supervised whole-slide image classification) with diverse classification data sets, including colorectal cancer histology, Patch Camelyon, prostate cancer detection, The Cancer Genome Atlas, and CIFAR-10, our task-specific self-supervised encoding approach consistently outperforms other convolutional neural network–based encoders. The better performances highlight the potential of task-specific attention-based self-supervised training in tailoring feature extraction for histopathology, indicating a shift from using pretrained models originating outside the histopathology domain. Our study supports the idea that task-specific self-supervised learning allows domain-specific feature extraction, encouraging a more focused analysis.

Introduction

Capitalizing on a wealth of labeled data, deep learning has showcased remarkable proficiency in analyzing medical images, even surpassing human performance.[1-3] However, the demanding and expensive nature of manual annotation has limited the availability of detailed annotations in the medical imaging field. This scarcity is especially prominent in histopathological whole-slide images (WSIs), which face challenges such as their large size (often gigapixel-scale), extensive heterogeneity, diverse cancer types, and variations in staining techniques.[4,5] WSIs encompass intricate biological structures, ranging from tiny cellular-level components such as subcellular vesicles and nuclear granules to larger tissue-level elements such as endothelia, epithelia, muscles, vessels, and glands. The immense size of these images makes labeling a vast and

* Corresponding author.
E-mail address: arahma34@jhu.edu (T. Rahman).

0893-3952/$ - see front matter
https://doi.org/10.1016/j.modpat.2024.100636

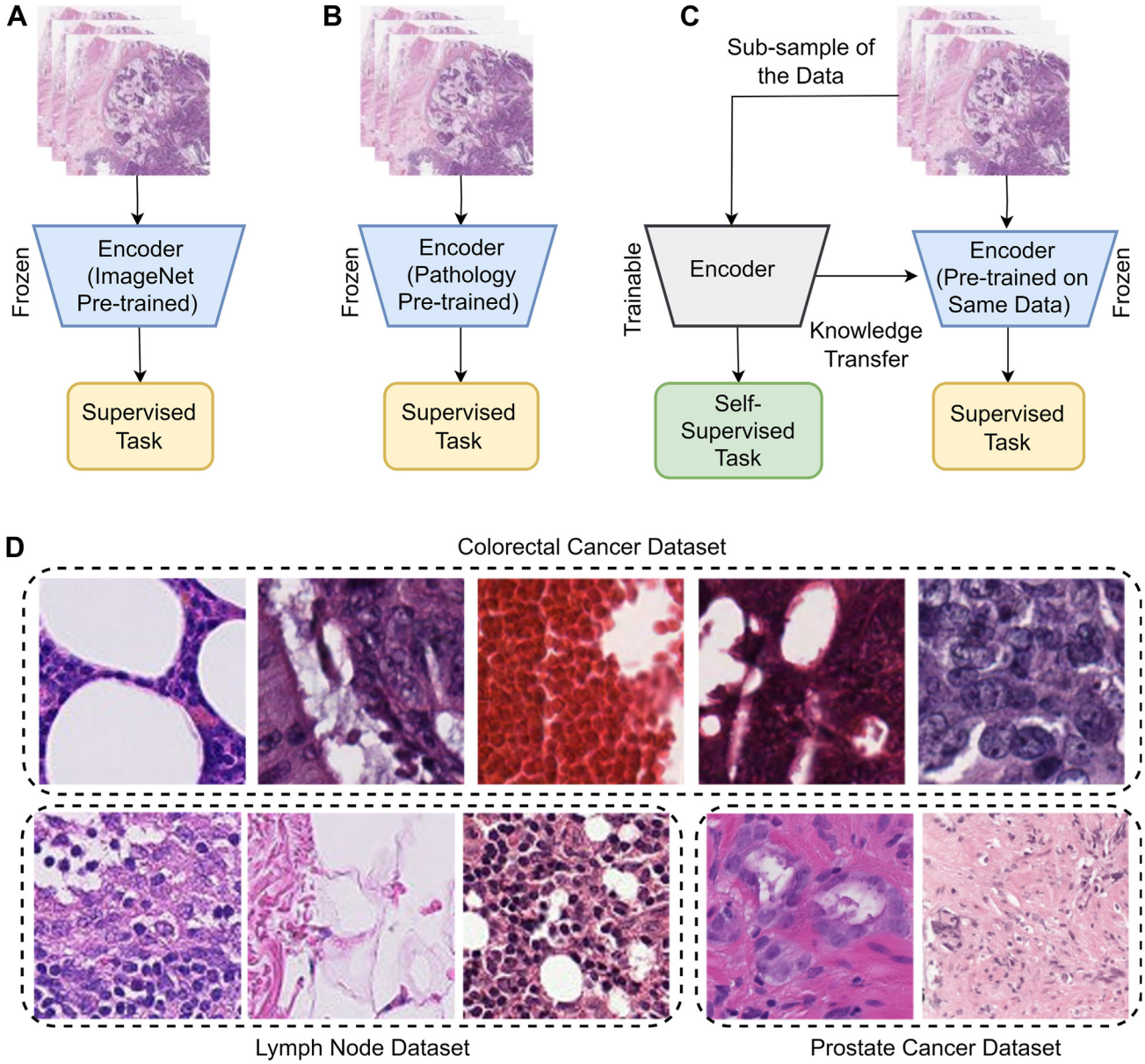

Figure 1.
Comparison among different pathology classification pipelines and our proposed TS-SSL, sample patch images of different histopathology data sets. (A) Ordinary classification pipeline that uses a frozen ImageNet pretrained encoder for patch feature extraction. There is a domain gap as the encoder is pretrained on images outside of the pathology domain. (B) Encoder is pretrained with pathology domain images. Although the domain shift problem is mitigated in this way, there is still a task-specific knowledge gap as it is not clear whether the pathology images used in pretraining have relevance to the downstream classification task. (C) The proposed TS-SSL pipeline in which a task-specific pretrained encoder is developed. It solves the domain shift problem and mitigates the knowledge gap in different tissue types, and (D) sample of diverse patch images for different tasks. TS-SSL, task-specific self-supervised learning.

complex task. Additionally, the diverse distribution of tissues makes it challenging to pinpoint small lesion regions within the entire WSI. The presence of multiple cancer types introduces various tissue patterns, making annotations more intricate. This complexity is further compounded by significant staining variations, leading to differences in color. Hence, there is an urgent need to develop an efficient tool for extracting features from unlabeled histopathological images. Such a tool would alleviate the burdensome task of annotation, fostering progress in digital pathology. This advancement could significantly aid pathologists by enabling faster and more accurate diagnoses.

In the domain of histopathology, the application of end-to-end training for image classification encounters significant computational demands due to the intricate nature and high resolution of histologic images. To mitigate this computational burden, a prevalent practice involves the freezing of the encoder, which optimizes computational resources by extracting essential features from histopathological samples as shown in Figure 1. These extracted features are then used as inputs for downstream tasks, such as classification, employing separate feed-forward networks. By segregating the feature extraction process and using distinct networks for subsequent tasks, computational efficiency is bolstered without compromising the accuracy and robustness of

classification models. Studies by Gecer et al[6], Hou et al[7], and Liu et al[8] underscore the significance of freezing encoders in extracting discriminative features and their subsequent use in downstream tasks, emphasizing the substantial impact on computational efficiency while maintaining classification efficacy in histopathology. Additionally, works by Bejnordi et al[9] and Litjens et al[10] further validate the efficacy of freezing encoders in optimizing computational resources for histopathological image analysis while preserving classification performance.

Moreover, to make histopathological images easier to work, one idea is to use transfer learning from big sets of labeled natural images such as ImageNet.[11] This method has shown that it can make a difference in how well we classify, measure, and break down different parts of these images with just a little bit of labeling.[12-15] But natural images and histopathological ones are different. Attributes or features such as textures and what they mean are not the same — like how objects and faces are unlike cells and tissues in medical images. Because of these differences, using only natural images might not help a lot. A better way could be training the learning system on specific medical data, but there are not enough labels for that. So, another way is to train the system without needing labels, letting it learn from the images themselves.

The remarkable achievements of self-supervised learning (SSL) techniques in computer vision have spurred their application in histopathological image analysis. Several studies have employed SSL to enhance the classification, regression, and segmentation of histopathological images, using existing contrastive learning (CL)-based SSL frameworks such as SimCLR and MoCo, or tailoring SSL tasks specifically for histopathology on convolutional neural network (CNN) backbones.[16-23] Although these approaches underscore the significance of SSL in histopathology, there are notable areas for improvement. Firstly, the bias in contrastive pairs defined in CL poses limitations for histopathological images, necessitating a tailored CL approach to enhance positive sample quality. Secondly, relying solely on CNN structures limits the extraction of both local and global features essential for analyzing histopathological images comprehensively. Additionally, the diversity of histopathological images poses a significant challenge for domain-based SSL, as illustrated by various task-specific patch images in Figure 1. Lastly, the relatively limited training data used in SSL training hinders the coverage of histopathological image diversity, highlighting the need for a universal SSL algorithm based on large scale, diverse data sets in the field.

To address these gaps, we propose a framework based on task-specific self-supervised learning, where subsamples of the same patch images are used for both self-supervised and supervised tasks to learn similar tissue types and cell adaptation, which also introduces a shallow feature extractor approach, this unique extractor is aimed to produce significantly more informative features tailored for specific tasks, concurrently enhancing processing speed. This study embarks on a comprehensive examination of the feature extraction process from digital pathology image tiles, aiming to elucidate the efficacy of self-supervised methods compared with models pretrained on ImageNet. Furthermore, we explored the complexities involved in patch classification and the subtleties of feature extraction, aiming to elucidate why self-supervised methods may present a more customized and resilient solution for the distinct challenges inherent in the landscape of digital pathology. By conducting a comprehensive evaluation of these methods, this research makes a valuable contribution to the progressing field of digital pathology. It assists in guiding the selection of optimal strategies for image classification and sets the stage for enhanced accuracy and efficiency in medical diagnosis.

Our key findings and contributions include the following:

1. We propose a task-specific self-supervised framework for histopathological image classification. We found that task-specific self-supervised pretraining significantly outperforms downstream classification tasks.
2. We also propose a shallow spatial-channel attention-based task-specific autoencoder for self-supervised pretraining, this unique pretrained encoder aimed to produce significantly more informative embeddings for specific tasks, concurrently enhancing processing speed.
3. Leveraging this design, our model demonstrates cutting-edge performance across 2 downstream classification tasks (such as patch classification and weakly supervised WSI classification), spanning 7 (4 for patch classification and 3 for weakly supervised WSI classification whole-slide label) histopathological data sets.
4. We demonstrate that self-supervised models are robust and generalize better than baselines trained using ImageNet, when pretrained on the subsample of the same downstream task data, without fine-tuning.

Notably, it exhibits heightened robustness and transferability compared with alternative SSL methods and both supervised and self-supervised ImageNet pretraining approaches. The efficacy displayed across these tasks signifies the potential of our proposed approach as a more efficient feature extractor strategy, poised to advance histopathological applications significantly.

Related Works

Transfer Learning for Histopathological Images

Transfer learning in histopathological images is a transformative approach revolutionizing the field by leveraging pre-existing knowledge from models trained on diverse data sets to enhance the analysis and interpretation of intricate tissue structures. In this methodology, pretrained models developed on large-scale data sets, such as ImageNet, serve as a foundation, enabling the extraction of meaningful features from histopathological images. By fine-tuning these models on domain-specific data sets, they can adapt to recognize nuanced patterns indicative of various diseases, facilitating tasks such as tumor identification, grading, and prognostication. A few recent works[16,17] exemplify the efficacy of transfer learning in histopathology, demonstrating its potential in cancer detection, classification of tissue types, and the identification of disease-specific morphologies. Cruz-Roa et al[16] illustrated the transferability of features learned from natural images to histopathological images, showing that models pretrained on ImageNet could be adapted to accurately classify breast cancer histology images. Similarly, Korbar et al[17] demonstrated that leveraging pretrained models improved the performance of classifiers for different tissue types in prostate cancer. Transfer learning not only expedites model convergence but also reduces the reliance on extensive annotated data sets, a significant advantage in histopathology where labeled data can be limited and resource-intensive to acquire. This approach democratizes access to state-of-the-art models, fostering advancements in computer-aided diagnosis and paving the way for more efficient and accurate histopathological image analysis techniques.

Self-Supervised Learning

Self-supervised learning has revolutionized image classification by allowing models to autonomously learn from unlabeled

data, a paradigm evident in numerous influential studies. Recent works[18-20] have significantly contributed to this domain. Doersch et al[18] introduced the concept of predicting relative patch positions within images as a pretext task, enabling models to grasp rich visual representations. Pathak et al[19] leveraged contextual information by predicting the color of missing patches, enhancing feature learning. Furthermore, Noroozi and Favaro[20] employed image rotations as a pretext task, highlighting the effectiveness of such self-supervised approaches in learning robust representations. The research efforts by Gidaris et al,[21] He et al,[22] and Zhang et al[23] further advanced SSL by exploring CL strategies, encouraging representations of similar images to be closer and those of dissimilar images to be farther apart in a learned feature space. Simultaneously, works by Chen et al,[24] Grill et al,[25] and Caron et al[26] introduced methodologies that learn representations by contrasting different augmentations of the same image, significantly enhancing the quality of learned features. These diverse studies collectively underscore the importance of SSL in image classification, showcasing its ability to learn powerful representations from unlabeled data, reducing the dependency on labeled data sets, and driving advancements in computer vision research.

Self-Supervised Learning in Histopathological Images

Using SSL in histopathological image classification has emerged as a promising avenue, empowering models to glean meaningful representations from unlabeled data and surmount challenges in limited annotated data sets. Pioneering works such as Hou et al[27] and Chen et al[28] have showcased the potential of self-supervised methods in this domain. Hou et al[27] introduced an SSL framework for histopathological image analysis by training models to predict the relative tissue order within pathology slides, enabling the extraction of salient features. Additionally, Chen et al[28] demonstrated the effectiveness of SSL by training models to predict tissue staining characteristics, contributing to improved classification accuracy in histopathological tasks. Recent research efforts,[29-31] have delved into CL strategies in histopathological images. Zhang et al[29] explored self-supervised CL to enhance feature representations, achieving robust performance in tissue classification tasks. Similarly, Wei et al[30] proposed a method leveraging self-supervised CL to encode tissue morphology and achieve state-of-the-art accuracy in histopathological image analysis. Furthermore, Xie et al[31] introduced an SSL framework to capture spatial context information for improved feature extraction in histopathological images. These diverse studies underscore the potential of SSL in histopathological image classification, demonstrating its efficacy in learning meaningful representations from unlabeled data. By harnessing SSL techniques, researchers aim to mitigate the need for extensive annotations while enhancing the interpretability and accuracy of models in analyzing intricate tissue structures. These advancements signal a transformative shift in histopathological image analysis, fostering more efficient and accurate diagnostic tools for medical practitioners and researchers.

Materials and Methods

Task-Specific Self-Supervised Learning Framework

Our proposed framework, task-specific SSL (TS-SSL), encompasses a multistep methodology tailored to optimize feature extraction for downstream supervised classification tasks. Initially, a Spatial-Channel Attention-based Autoencoder (scAE) was trained using a subset of the tile data set, focusing on learning discriminative representations. Figure 2 illustrates the subsequent utilization of the encoder component derived from the pretrained scAE to extract task-specific features. This process involves integrating attention mechanisms within the autoencoder, enabling the model to selectively emphasize relevant features during training. Mathematically, scAE was trained for a pretext task using a subset X' from the downstream data set where X represents the entire downstream data set, focusing on learning discriminative representations. The scAE is trained by minimizing the reconstruction loss, represented as:

$$L_{\text{reconstruction}} = \|X' - D_{\text{scAE}}(E_{\text{scAE}}(X'))\|_2^2 \quad (1)$$

Here, E_{scAE} and D_{scAE} represent the encoder and decoder components of the scAE, respectively. X' undergoes encoding and decoding, aiming to reconstruct the input data, minimizing the difference between the original and reconstructed data. Following the scAE training, the pretrained encoder (E_{scAE}) extracts task-specific features f_{TS} from the entire data set, denoted as $f_{\text{TS}} = E_{\text{scAE}}(X)$. The integration of attention mechanisms during scAE training enables the model to emphasize pertinent features. Then, we fed the feature embeddings to a multilayer perceptron (MLP) block for the final downstream patch classification task and used attention-based multiple instances learning block for the weakly supervised WSI classification task.

Spatial-Channel Attention Autoencoder

The autoencoder consists of an encoder and a symmetrical decoder part. Two main blocks of the autoencoder, namely, spatial attention and channel attention, take the feature map after the convolutional block (it consists of 3 $\times$ 3 convolution, batch normalization, and rectified linear unit activation) as an input and then produce 2 new feature maps of equal size, shown in Figure 2. After concatenating the new feature maps and the original one into a bigger feature map, which is applied to another convolutional block.

Spatial Attention

The Spatial Attention block (Fig. 2) integrates a nonlocal operation self-attention mechanism to capture context relations among local areas in the input feature map. Widely used in natural language processing, recent studies[32-34] show its efficacy in enhancing semantic segmentation with CNNs. Operating on the original feature map, this block employs 3 parallel convolutional layers to learn diverse context relations. Reshaping, activation, max pooling, and matrix multiplication handle the intermediate layer results. The final convolutional layer produces a new feature map encapsulating context relation between local features extracted, contributing to the improvement in feature representation. In Figure 2, the spatial attention block processes the original feature map $\in R^{m \times n \times c}$. Each feature in M encapsulates local details from the input image. Feature map f is derived via convolution and reshaping, which is mentioned as β, mathematically defined as, $f = \beta(M)$.

The Reshape function alters the matrix's shape while preserving its original data, and *Conv* represents a 1 $\times$ 1 convolution operation. Additionally, feature maps Q and V are obtained through convolution, reshaping, and max pooling. Contextual relationships among features are determined through the nonlocal operation in the embedded Gaussian version.[1] By applying the

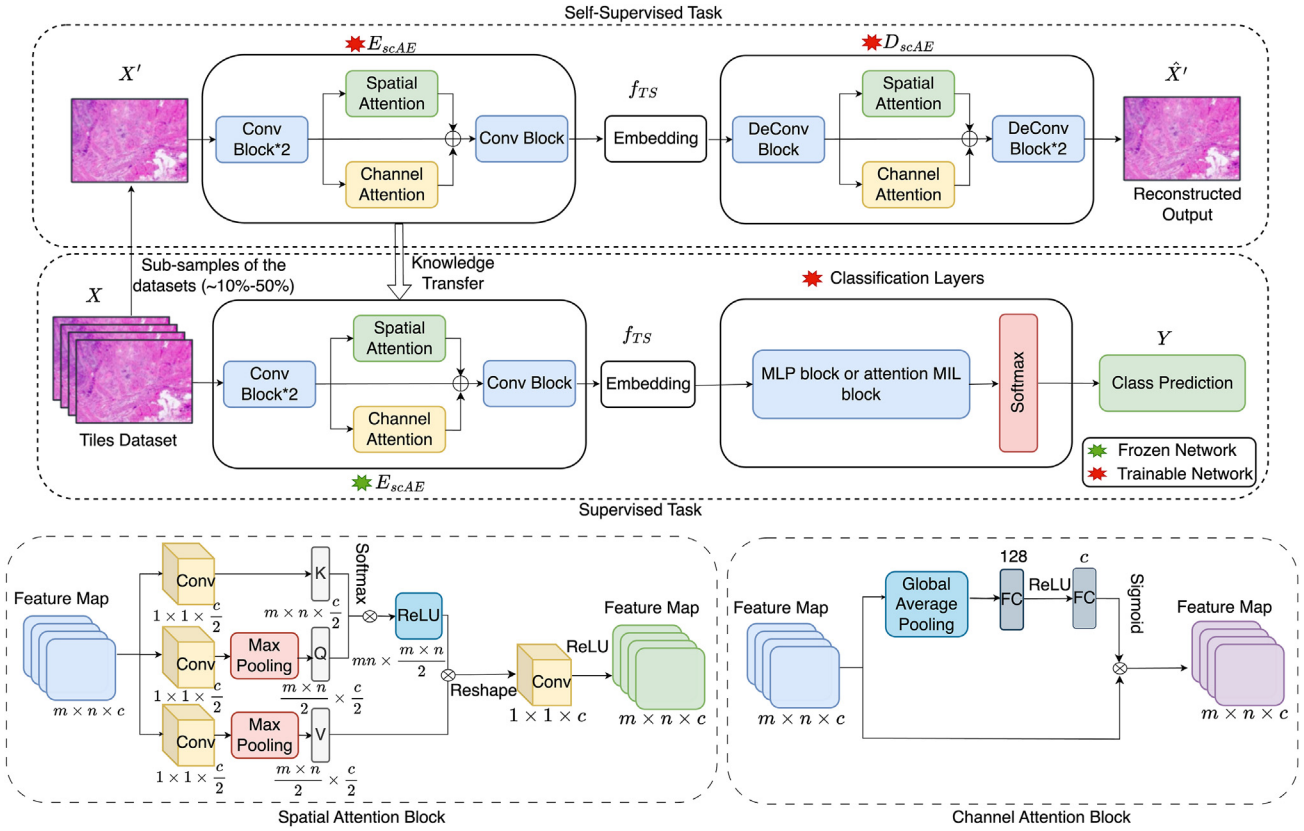

Figure 2.
Our TS-SSL framework. An illustration of our self-supervised pretraining for histopathological image classification. When subsample of same patch images are used for both self-supervised task and downstream supervised task to learn same tissue and cell adaptation. Two main blocks of the scAE, namely, spatial attention and channel attention block. TS-SSL, task-specific self-supervised learning

Softmax function to the matrix $KQ \in R^{m\ n\ x\ \left[\frac{mn}{2}\right]}$ and multiplying it by V, the weights for spatial attention, α between different local features are derived:

$$\alpha = \text{Softmax}\left(KQ^T\right) V \tag{2}$$

Finally, a new feature map $M' \in R^{m\ x\ n\ x\ c}$ with contextual relation information is generated using rectified linear unit activation after a 1×1 convolution operation, mathematically defined as, $M' = \beta(\alpha)$.

Channel Attention

The Channel Attention block (Fig. 2) is designed to capture channel-wise attention by learning weights associated with individual channels. Research studies[35,36] affirm that feature map channels from CNNs carry varying degrees of semantic information. Using this insight improves image classification outcomes. Initially, channel weights in Figure 2 of the original feature map $M \in R^{m\ x\ n\ x\ c}$ are initialized via global average pooling, Φ. Following this, 2 consecutive fully connected layers are trained to determine new weights for the original features. Ultimately, a new feature map F' is generated by combining these new weights with the original features, enhancing the model's capacity to capture essential information and thereby improving image classification, mathematically defined as,

$$F' = M \circledast \sigma(\Phi(M)) \tag{3}$$

Classification Block

Multilayer Perceptron Block

In this study, we used an MLP block after the encoder part for the final downstream patch classification task. The extracted features were subsequently fed into a MLP architecture comprising 2 dense layers and 1 dropout layer, followed by a final classification layer. Mathematically, $F_{\text{task-specific}}$ denote the extracted features obtained from the pretrained encoder. The extracted features are then input into the MLP block. The output of the first dense layer and incorporating a dropout regularization technique to prevent overfitting, denoted as $H_{1\text{dropout}} = \text{Dropout}(\sigma(W_1 F_{TS} + b_1))$. W_1 represents the weights, b_1 denotes the biases, and σ signifies the activation function used. The final classification layer computes the output probabilities for classification tasks, given by,

$$Y = \text{softmax}\left(W_2 H_{1\text{dropout}} + b_2\right) \tag{4}$$

Where W_2 and b_2 represent the weights and biases of the classification layer, respectively.

Attention Multiple Instance Learning Block

In weakly supervised WSI classification tasks, extracted features underwent a transformative stage through an attention-based multiple instance learning (MIL) block, designed to imbue the model with the capability to discern essential patterns within the feature set.[37] The architecture of the MIL block incorporated an attention mechanism that selectively emphasized crucial features by assigning attention weights to each feature. Mathematically, the attention mechanism can be formulated as follows: let $F_{TS} = \{h_1, h_2, \ldots\ldots\ldots, h_K\}$ represent the bag of K embeddings extracted by the pretrained encoder. The attention MIL pooling,

$$M = \sum_{k=1}^{K} A_k h_k \tag{5}$$

Where,

$$A_k = \frac{\exp\left\{w_a{}^T \tanh\left(V h_k{}^T\right)\right\}}{\sum_{j=1}^{K} \exp\left\{w_a{}^T \tanh\left(V h_j{}^T\right)\right\}}$$

Where w_a and V are parameters. Additionally, we employed the hyperbolic tangent function (tanh) as a nonlinear element-wise operation, encompassing both negative and positive values to ensure appropriate gradient flow. The attended features M were derived by computing the element-wise multiplication of the attention weights and the extracted features. These attended features M were then channeled into the subsequent final classification layer, denoted as, $Y = \text{softmax}(W_c M + b_c)$. Where W_c and b_c symbolize the weights and biases of the classification layer, respectively. This attention-driven MIL strategy facilitated enhanced discrimination of salient information, fostering improved performance in classification tasks by enabling the model to focus on pivotal features.

Task Data set and Implementation Setup

Data Set. In this study, we included 3 pathology domain classification tasks and one out of pathology domain in which tile-level labels were available: namely, identification of metastatic breast cancer in lymph nodes (Pcam),[9] prostate cancer detection (PANDA),[38] colon cancer tissue compartment labeling (CRC),[39] and out of domain data set (CIFAR-10).[40] Furthermore, at the whole slide-level classification in the pathology domain we examined the Cancer Genome Atlas (TCGA)-non-small cell lung cancer (NSCLC) subtyping (adenocarcinoma vs squamous cell carcinoma),[41] TCGA-breast cancer (BRCA) subtyping (ductal vs lobular carcinoma),[41] and PANDA (prostate cancer detection at whole-slide level).[41] All these data sets are introduced below.

Patch Camelyon. The Patch Camelyon (PCam) data set comprises 327,680 histopathology images at 96 × 96 resolution, extracted from lymph node sections. Each image exhibits tissue structures, labeled for cancer presence, enabling binary classification. This data set is divided into a training set comprising 262,144 images and a separate test set containing 65,536 images. PCam provides a balanced distribution between positive (cancerous) and negative (benign) samples. With annotations indicating tumor regions, this data set serves as a pivotal resource for training and evaluating machine learning models in cancer detection.

Prostate Cancer Detection. In this data set, WSIs of prostate biopsies were retrospectively collected from 6 different sites for algorithm development, tuning, and independent validation. In this study, we preprocessed the data set for 2 downstream tasks. For the WSI classification task, 1925 slides were categorized as benign, whereas 4932 were classified as cancerous tissue. In the patch classification task, we divided each slide into patches measuring 256 × 256 pixels. Of these patches, 98,438 were classified as benign, whereas 285,565 were labeled as cancerous for binary classification.

Colorectal Cancer. Hematoxylin and eosin–stained CRC tissue slides were obtained from the pathology archive at the University Medical Center Mannheim (Heidelberg University, Mannheim, Germany), proposed for colorectal classification task. It is composed of 5000 patches with a size of 150 × 150 pixels (74 × 74 microns) and covers 8 different tissue types (625 patches for each type), including epithelium, simple stroma, complex stroma, lymphoid follicles, debris, mucosal glands, adipose and background region of interest with no tissue.

CIFAR-10. The CIFAR-10 data set is a widely used benchmark in machine learning, particularly in image classification tasks. It consists of 60,000 color images, each sized 32 × 32 pixels, distributed across 10 different classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The data set is split into 50,000 training images and 10,000 test images, with each class containing 6000 images.

The Cancer Genome Atlas. TCGA is a public large-scale multimodal data set, which contains genome, epigenome, transcriptome, and image data. This work only considers the image data (frozen and formalin-fixed paraffin-embedded [FFPE] slides), which includes a total of 30,072 WSIs covering over 25 anatomical sites and over 32 cancer subtypes. For each WSI, a primary diagnosis is provided for the entire WSI but no detailed annotations. Three hundred nine WSIs are removed due to the lack of magnification information. In this study, we used 2 cancer subtype data sets.

The Cancer Genome Atlas-Non-Small Cell Lung Cancer. TCGA-NSCLC is collected from the TCGA data set for 2 types of lung cancer classification: lung squamous cell carcinoma and lung adenocarcinoma, which consists of a total of 993 FFPE WSIs (507 WSIs with lung adenocarcinoma and 486 WSIs with lung squamous cell carcinoma).

The Cancer Genome Atlas-Breast Cancer. TCGA-BRCA is collected from the TCGA data set for 2 types of breast carcinoma classification: invasive lobular carcinoma and invasive ductal carcinoma, which consists of a total of 938 FFPE WSIs (772 WSIs with invasive ductal carcinoma and 166 WSIs with invasive lobular carcinoma).

Pretraining Setup

To assess the effectiveness of self-supervised pretraining, we investigated ResNet-50 (pretrained on ImageNet), ResNet-50 (random initialized), DenseNet-121 (pretrained on ImageNet), DenseNet-121 (random initialized), ResNet based foundation model (pretrained on histopathology images),[42] KimiaNet (pretrained on histopathology images),[43] PLIP ,[44] CTransPath,[44] CONCH,[45] and our TS-SSL architectures as base encoder networks. Following this, an MLP block (for patch classification) or attention MIL block (for weakly supervised WSI classification) was used to map the output of the encoder, which is used for final class prediction. For the TS-SSL framework, we trained scAE for the subsample of each data set and then used the encoder part for the downstream task of the same

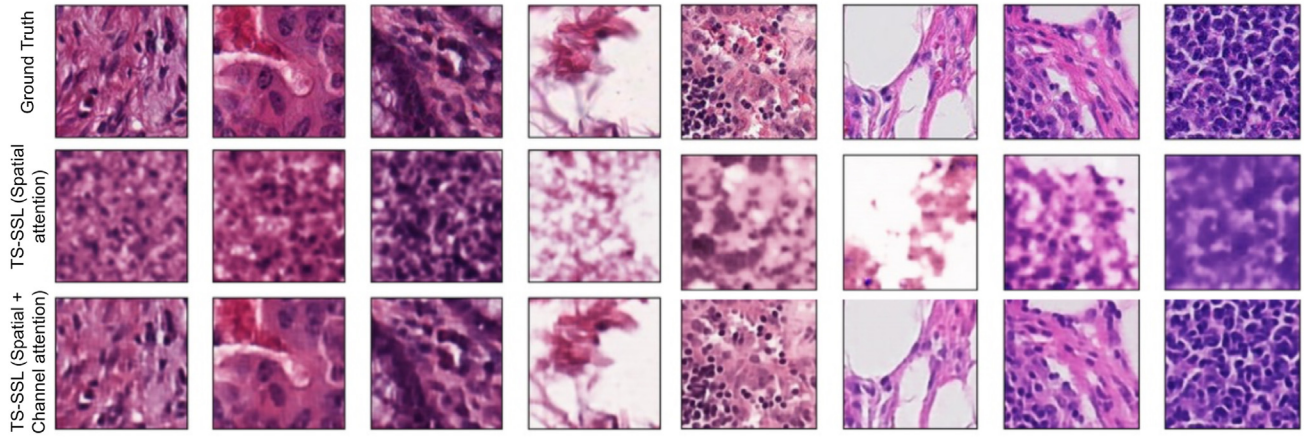

Figure 3.
Task-specific self-supervised learning (TS-SSL) reconstruction. First row: ground truth patch image. Second row: reconstructed patch images from TS-SSL-based autoencoder using only spatial attention. Third row: reconstructed images from TS-SSL-based autoencoder using both spatial and channel attention.

data set. Moreover, we trained the autoencoder for up to 1500 steps and found that the batch size of 256 and learning rate of 0.003 works well in this setting. For the downstream patch classification tasks, we trained the models for up to 5000 steps with a batch size of 128 and a learning rate of 0.001. We trained the models for up to 2500 steps with a batch size of 128 and a learning rate of 0.01 for WSI classification in the downstream task.

Evaluation Methodology

After determining optimal hyperparameters for fine-tuning a specific data set, we selected the model based on its performance on the validation set. Subsequently, we evaluated the chosen model multiple times (5 iterations for patch classification and 3 iterations for WSI classification) on the test set to report task performance. Our primary metrics for both tasks were top-1 accuracy and area under the curve (AUC) score. Given the multilabel setup, we calculated the top-1 accuracy averaged across predictions for all target classes following He et al.[22]

Results

In this section, we delve into whether using a self-supervised pretraining approach with a task-specific-attention autoencoder leads to enhanced performance in models fine-tuned end-to-end for specific histopathological image classification tasks. Initially, we examined the optimal choice of pretraining weights tailored for these classification tasks. Following this, we assessed the advantages of our proposed framework, which employed task-specific attention in a self-supervised manner, for both the patch and weakly supervised WSI classification tasks and compared its efficacy against baseline methods and cutting-edge approaches in supervised pretraining. Lastly, we investigated the efficiency, in terms of both labeling and computational resources, of models trained via self-supervision in the context of histopathological image classification.

Task-Specific Self-Supervised Learning Based scAE Reconstruction

We present the reconstruction outcomes achieved by the spatial attention-based autoencoder (scAE) in Figure 3. The first row showcases the ground truth patch image to provide a baseline for comparison, the second row, we observed the reconstructed patch images generated by the TS-SSL-based autoencoder using solely spatial attention. The third row shows reconstructed images from the same autoencoder but incorporating both spatial and channel attention. The compelling results illustrate the capability of the model to recover information from both global and local contexts. Notably, the primary objective of the model is not centered on producing high-fidelity reconstructions but rather on enhancing performance on downstream tasks. We hypothesize that in medical imaging, where local areas are intricately linked to their physiological surroundings, contextual information plays a crucial role in accurately reconstructing image patches. Proficiency of the reconstruction model in preserving such information holds promise for its application in medical image analysis and other contexts where contextual understanding is paramount.

Patch Classification

In Table 1,[43-45] we have tabulated the outcomes of patch classification using our method and conducted a comparative analysis with various established techniques commonly employed in histopathological image analysis. Furthermore, our method underwent rigorous evaluation across 3 distinct histopathological data sets. Notably, our approach exhibited superior performance across all data sets, surpassing the widely used ResNet50 with ImageNet pretraining by a notable margin of ~12%. Of particular interest is the observation that ResNet50 with randomly initialized weights can achieve higher accuracy than its counterpart initialized with ImageNet weights, specifically in the case of CRC and Patch Camelyon data sets.

We trained and evaluated all methods by using 5-fold cross-validation across PANDA, CRC, PCam, and CIFAR-10 data sets. Our TS-SSL-based patch classification consistently outperformed all the other methods. We achieved accuracies of 92.11%, 86.6%, 89.51%, and 68.22%, for PANDA, CRC, PCam, and CIFAR-10 data sets, respectively. Our approach improved the performance from approximately +1% to +12% over some of the currently available foundational SSL models and the more conventional ImageNet pretraining approach. These findings support the assertion that the proposed TS-SSL method is a performant and efficient

Table 1
Comparison of different methods with our proposed method for patch classification

Architecture	Model state	PANDA		CRC		PCam		CIFAR-10	
		Top1% accuracy	AUC	Top1% accuracy	AUC	Top1% accuracy	AUC	Top1% accuracy	AUC
ResNet50	Random initialized	81.5	82.14	72.81	74.38	79.96	80.58	47.02	47.39
ResNet50	ImageNet pretrained	83.56	85.21	60.5	63.97	76.35	76.95	64.97	65.48
DenseNet121	Random initialized	78.84	79.45	58.31	58.76	82.45	83.09	59.85	60.32
DenseNet121	ImageNet pretrained	83.88	84.53	58.63	59.09	85.6	86.27	66.56	67.08
ResNet-based Foundation model[42]	Pathology pretrained	87.65	88.33	71.8	73.36	87.58	88.26	-	-
KimiaNet (Foundation model)[43]	Pathology pretrained	88.24	89.93	71.95	72.51	88.83	89.52	-	-
PLIP[44]	Pathology pretrained	90.04	91.56	75.58	78.45	88.91	90.11	-	-
CONCH[45]	Pathology pretrained	90.88	92.25	82.56	83.33	88.56	89.45	-	-
CTransPath[44]	Pathology pretrained	88.68	90.12	73.08	74.86	87.89	88.78	-	-
TS-SSL (ours)	Task-specific pretrained	92.11	93.83	86.6	87.28	89.51	91.21	68.22	71.75

AUC, area under the curve; CIFAR-10, out of domain data set; CRC, colorectal cancer; PCam, identification of metastatic breast cancer in lymph nodes; TS-SSL, task-specific self-supervised learning.

approach to learning discriminative features for a given specific supervised learning task.

Whole-Slide Image Classification

Our study next examined the discriminative potential of the proposed task-specific self-supervised representation learning in the context of weakly supervised classification experiments across 3 WSI-level data sets: TCGA-NSCLC, TCGA-BRCA, and PANDA. This investigation also involved a meticulous comparison between weakly supervised classification outcomes based on TS-SSL-pretrained features and the current state-of-the-art methods, detailed in Table 2.[42-45] At the WSI level, a key challenge is the nature of the weakly supervised in which only global annotations (slide level) labels are available and patch/regional labels are not available. Existing algorithms for WSI classification typically entail 2 key phases: (1) extracting features from patches cropped from WSIs and (2) aggregating these patch features. Various feature extraction methods were explored, including ImageNet pretraining, SSL pretraining, and our novel TS-SSL pretraining. Additionally, aggregation algorithms leveraging attention-based pooling for feature fusion were considered. In our method, the feature extractor adopted our TS-SSL-pretrained model, whereas the feature aggregator directly used MIL (attention-based pooling) within the base classification block.[37,46]

Employing 5-fold cross-validation across the TCGA-NSCLC, TCGA-BRCA, and PANDA datasets revealed compelling findings. Our TS-SSL-based weakly supervised classification consistently outperformed all other methodologies, achieving accuracies of 86.56%, 85.57%, and 97.77% for TCGA-NSCLC, TCGA-BRCA, and PANDA, respectively. This demonstrated an approximate enhancement of +2% and +5% over the foundational SSL method and ImageNet pretrained features. These outcomes further substantiate the robust feature learning prowess encapsulated within our TS-SSL-based framework.

Efficient Pretraining of TS-SSL Via Random Sub-Sampling of WSIs

We investigated the number of patches required from the slide-level data set to effectively train an excellent encoder. Our findings shown in Table 3 highlight the results of WSI classification. We randomly selected varying percentages—1%, 5%, 10%, 50%, and 100%—of patches for TS-SSL pretraining. Subsequently, we evaluated the performance of our TS-SSL-based method for weakly supervised WSI classification. To our surprise, the TS-SSL-pretrained model consistently outperformed, even when using only 1% to 10% of the patch data for the initial task. This unexpected outcome suggests that merely 10% of the patches sufficed to prepare our TS-SSL method for the subsequent task, showcasing its remarkable efficiency in performing well with minimal data. The capability to achieve remarkable results with just 10% of the

Table 2
Comparison of different methods with our proposed method for weakly supervised WSI classification

Encoder	Pretrained data set	TCGA-NSCLC		TCGA-BRCA		PANDA	
		Top1% accuracy	AUC	Top1% accuracy	AUC	Top1% accuracy	AUC
ResNet50	Random initialized	78.91	79.6	76.75	78.43	93.72	95.54
ResNet50	ImageNet pretrained	81.09	83.8	78.89	79.58	94.15	94.98
DenseNet121	Random initialized	76.61	79.28	78.03	80.72	95.23	97.07
DenseNet121	ImageNet pretrained	80.01	80.71	79.1	79.8	94.88	96.71
ResNet-based foundation model[42]	Pathology pretrained	84.56	85.3	82.08	82.8	96.83	97.68
KimiaNet (Foundation model)[43]	Pathology pretrained	85.07	85.82	83.15	83.88	96.26	97.11
PLIP[44]	Pathology pretrained	85.45	86.08	84.65	85.86	96.88	97.43
CONCH[45]	Pathology pretrained	85.76	86.88	84.32	85.12	96.65	97.08
CTransPath[44]	Pathology pretrained	84.88	85.78	81.5	83.68	95.53	96.92
TS-SSL (Ours)[a]	Task-specific pretrained	86.56	88.43	85.51	87.37	97.77	98.75

AUC, area under the curve; TCGA-NSCLC, The Cancer Genome Atlas-non−small cell lung cancer; TCGA-BRCA, The Cancer Genome Atlas-breast cancer; TS-SSL, task-specific self-supervised learning; WSI, whole-slide images.
[a] Based on 10% subsampling training (see Table 3).

Table 3
Performance of our method for TS-SSL training with different percentages of patch images

Method	#Percentage of patches for training	TCGA-NSCLC		TCGA-BRCA		PANDA	
		Top1% accuracy	AUC	Top1% accuracy	AUC	Top1% accuracy	AUC
TS-SSL	1	84.78	85.63	83.65	84.49	96.65	97.62
	5	85.4	86.25	85.12	85.97	97.06	98.03
	10	86.56	88.43	85.51	87.37	97.77	98.75
	50	86.22	87.18	84.65	85.50	97.85	98.83
	100	85.65	86.51	85.12	85.97	97.65	98.63

AUC, area under the curve; PANDA, prostate cancer detection; TCGA-NSCLC, The Cancer Genome Atlas-non−small cell lung cancer; TCGA-BRCA, The Cancer Genome Atlas-breast cancer; TS-SSL, task-specific self-supervised learning.

patches prepared for training underscores the crucial flexibility of our method, particularly in situations where acquiring extensive data proves challenging or expensive. This reveals the adaptability and effectiveness of our method, demonstrating its potential to perform effectively even in scenarios with limited data, making it versatile across different situations.

Furthermore, we compared the effectiveness of our approach using only spatial attention and those employing both spatial-channel attention. These findings are summarized in Supplementary Figure S1A,B. The results indicate that both approaches using spatial-channel attention outperformed the sole use of spatial attention-based autoencoders with ~2.5% accuracy. We also cf the AUC performance, which shows an average of ~3% improvement using both attentions.

Visualization

We explored *t*-distributed stochastic neighbor embedding visualization for encoded feature representations of the CRC data set using 3 different pretrained encoders: random initialized, ImageNet pretrained, pathology pretrained, and task-specific pretrained. We also calculated the Silhouette coefficient (SC) and Davies−Bouldin index (DBI) metrics for these representations through K-means clustering,[47-49] shown in Supplementary Table S1. The SC evaluates how well separated the clusters are, with higher values indicating better-defined clusters. The DBI measures cluster compactness and separation, with lower values suggesting better clustering performance. These metrics were chosen because they provide complementary views of clustering quality: although SC focuses on individual point cohesion and separation, DBI looks at the overall cluster structure. We have now included confidence intervals to better represent the variability and reliability of our clustering performance metrics. The confidence intervals and clustering metrics provided are averages over 10 runs of the K-means algorithm, each with a different random initialization. Including confidence intervals and averaging over multiple runs allows us to demonstrate that our clustering results are both consistent and statistically significant. This robustness adds credibility to the claim that the model representations learned by our method yield better clustering results compared with baseline methods.

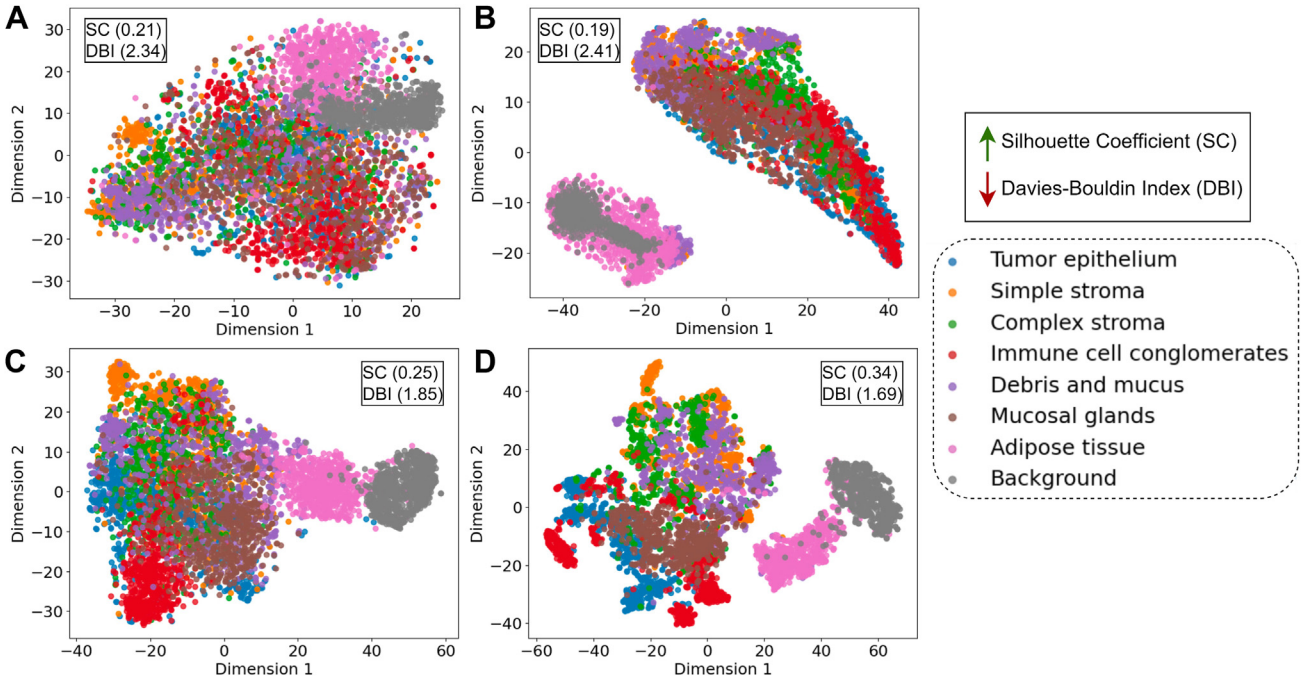

Figure 4.
The *t*-distributed stochastic neighbor embedding visualization and radar plot for clustering performance score. The results of encoded feature representations from colorectal cancer (CRC) data set using freezing encoders. (A) random initialized, (B) ImageNet pretrained, (C) pathology pretrained (PLIP), (D) our task-specific pretrained.

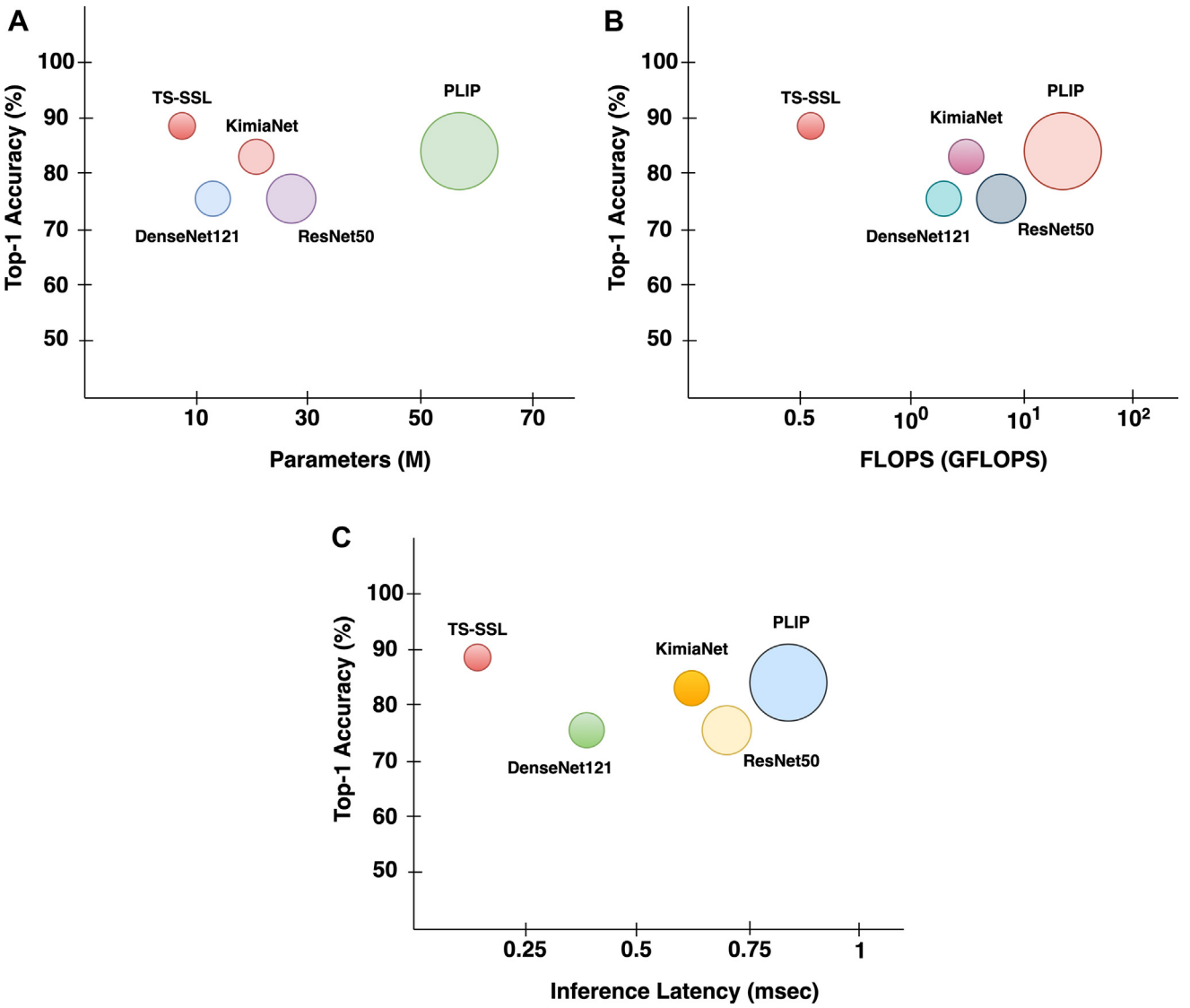

Figure 5.
Comparison of computational complexity. Comparison of the parameters and computational expenses of TS-SSL with other networks (A) Top-1% accuracy vs FLOPS, (B) Top-1% accuracy vs model parameters, and (C) Top-1% accuracy vs model's inference latency. TS-SSL, task-specific self-supervised learning.

Both random initialization and ImageNet pretrained encoders exhibited a minimal degree of alignment with the known labels (Figure 4A, B). There is some degree of segregation of the "background" and "adipose" classes; however, this is likely simply driven by the high degree of background pixels in both of these classes relative to the others. In contrast, the pathology pretrained and our proposed task-specific pretrained encoders resulted in representations that better aligned with the known labels. Figure 4D specifically showcases the CRC data set's feature representation distributions via the task-specific pretrained encoder, revealing more distinct clusters for different classes. These clusters exhibited morphology characteristics of their respective subtypes, indicating the efficacy of task-specific self-supervised learning. We can also see that our approach exhibits better performance metrics compared with other conventional approaches in Supplementary Table S1. This approach enables the encoder to assign discernible features to class patterns, aiding the classification model in predicting classes based on the generated feature map. These results strongly suggest that the learned features accurately represent cancer morphology. This quality significantly contributes to the superior performance of our system compared with other approaches.

Computational Comparison

Our less-complex TS-SSL models consistently achieve a reduction in parameters and floating point operations per second (FLOPS) by a significant margin, which is up to 3.2 times fewer parameters and up to 7.5 times fewer FLOPS compared with the existing ResNet50 model for end-to-end training. Moreover, it is up to 73 times fewer parameters and up to 40 times fewer FLOPS compared with the existing ResNet50 model for only the encoder part. Figure 5A, B provide a comprehensive comparison of TS-SSL's parameter count with that of other networks, showcasing its superior efficiency in terms of model parameters when contrasted with alternative architectures, which are also illustrated in Supplementary Table S2. Additionally, we computed the inference

latency using the TS-SSL-based encoder in our proposed framework. We contrasted it with the inference latencies of other frequently employed encoders for histopathological feature extraction. As illustrated in Figure 5C, the comparison of various encoders' inference latency reveals that our approach is ~4 times faster than the commonly used ResNet50 pretrained encoder for extracting patch features, also provided in Supplementary Table S3. This highlights the efficiency of our proposed framework in terms of processing speed, demonstrating a notable advantage in histopathological image analysis tasks.

Discussion

The proposed TS-SSL framework is an effective method to train and use scAE for feature extraction, showcasing the synergy between attention mechanisms and self-supervised learning. The spatial and channel attention blocks within the scAE play crucial roles. Spatial attention introduces a nonlocal operation self-attention mechanism, enhancing semantic segmentation by capturing contextual relations among local features. Concurrently, channel attention focuses on channel-wise semantic information, significantly improving image classification outcomes. These attention blocks contribute to our framework's superior performance by enhancing feature representations that are crucial for accurate classification.

Across a variety of data sets, including PCam, CRC, CIFAR-10, TCGA-NSCLC, TCGA-BRCA, and PANDA, our method consistently outperforms established techniques in 2 downstream classification tasks. It exhibits superior performance compared with common approaches such as ResNet50 with ImageNet pretraining, emphasizing the discriminative feature learning capabilities embedded in our methodology. Specifically, our TS-SSL-based patch classification achieves remarkable accuracy, surpassing all other methods with percentages of 92.11%, 86.6%, 89.51%, and 68.22% on PANDA, CRC, PCam, and CIFAR-10 data sets. Notably, our approach enhances performance by +1% to +12% over foundational SSL models and ImageNet pretraining, showcasing the effectiveness of TS-SSL in learning discriminative features for specific supervised tasks. Additionally, our TS-SSL-based weakly supervised classification consistently outperforms alternative methodologies, attaining accuracies of 86.56%, 85.57%, and 97.77% for TCGA-NSCLC, TCGA-BRCA, and PANDA, respectively. This demonstrates a substantial +2% to +5% improvement over foundational SSL and ImageNet pretraining. The results from both patch and weakly supervised WSI classification highlight the adaptability and generalization of our framework across various histopathological data sets. Our subsampling-based pretraining of TS-SSL underscores the efficiency of the proposed approach. The ability of the model to achieve superior performance with minimal percentages of patch data demonstrates its robust feature learning even in resource-constrained scenarios. Furthermore, the computational efficiency of our TS-SSL models, significantly reducing parameters and FLOPS compared with prevalent architectures such as ResNet50, ensures an efficient yet high-performing solution for histopathological image analysis. In essence, the TS-SSL framework not only excels in accuracy but also showcases computational efficiency, making it a promising solution for real-world medical image analysis applications. The effective amalgamation of attention mechanisms and SSL in our approach demonstrates its potential to address challenges in histopathological image classification with improved accuracy and resource utilization.

In conclusion, our study provides vital insights into feature extraction in digital pathology. We explored tile-encoding strategies, revealing the most robust approach and shedding light on the extent of "transfer learning." The evaluation of neural network architectures, including our task-based self-supervised feature extraction method, proved effective for histopathology's intricacies. This approach consistently outperformed CNN-based encoders, emphasizing its potential for patch classification. Self-supervised learning promises to revolutionize digital pathology workflows, particularly when aggregating features across encoded tiles, enhancing accuracy and efficiency. In summary, SSL is a valuable alternative to ImageNet pretraining, optimizing feature extraction for precise medical diagnostics and research in digital pathology.

Author Contributions

T.R., A.B., and R.C. performed conceptualization; T.R. performed data processing, formal analysis, and investigation; T.R., A.B., and R.C. contributed to methodology, writing – original draft preparation and review and editing. All authors have read and agreed to the published version of the manuscript.

Data Availability

All data sets are publicly available. TCGA-NSCLC and TCGA-BRCA data in this study are obtained from TCGA via the Genomic Data Commons Data Portal (https://gdc.cancer.gov). The code will be released to a public GitHub repository upon acceptance of the manuscript.

Funding

The author(s) received no specific funding for this work.

Declaration of Competing Interest

The authors have no competing interests to disclose.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2024.100636.

References

1. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719–731. https://doi.org/10.1038/s41551-018-0305-z
2. Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mach Intell*. 2019;1(5):236–245. https://doi.org/10.1038/s42256-019-0052-1
3. Zhang R, Isola P, Efros AA. Colorful Image Colorization. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science; vol 9907. Springer; 2016:649–666. https://doi.org/10.1007/978-3-319-46487-9_40
4. Rashid R, Chen Y-A, Hoffer J, et al. Narrative online guides for the interpretation of digital-pathology images and tissue-ATLAS data. *Nat Biomed Eng*. 2022;6(5):515–526. https://doi.org/10.1038/s41551-021-00789-8
5. Javed S, Mahmood A, Fraz MM, et al. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Med Image Anal*. 2020;63:101696. https://doi.org/10.1016/j.media.2020.101696
6. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-based deep neural networks for detection of cancerous and precancerous

esophagus tissue on histopathological slides. *JAMA Netw Open*. 2019;2(11): e1914645. https://doi.org/10.1001/jamanetworkopen.2019.14645
7. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016;2016:2424–2433. https://doi.org/10.1109/CVPR.2016.266
8. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. *Preprint*. Published online on March 3, 2017. bioRxiv 1703.02442. https://doi.org/10.48550/arXiv.1703.02442
9. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585
10. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005
11. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
12. Talo M. Automated classification of histopathology images using transfer learning. *Artif Intell Med*. 2019;101:101743. https://doi.org/10.1016/j.artmed.2019.101743
13. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal*. 2021;67:101813. https://doi.org/10.1016/j.media.2020.101813
14. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Med*. 2020;26(6):900–908. https://doi.org/10.1038/s41591-020-0842-3
15. Mormont R, Geurts P, Marée R. Multi-task pre-training of deep neural networks for digital pathology. *IEEE J Biomed Health Inform*. 2020;25(2): 412–421. https://doi.org/10.1109/JBHI.2020.2992878
16. Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci Rep*. 2017;7:46450. https://doi.org/10.1038/srep46450
17. Korbar B, Olofson AM, Miraflor AP, et al. Deep learning for classification of colorectal polyps on whole-slide images. *J Pathol Inform*. 2017;8:30. https://doi.org/10.4103/jpi.jpi_34_17
18. Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. *Conf Comput Vis*. 2015:1422–1430. https://doi.org/10.1109/ICCV.2015.167
19. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: feature learning by inpainting. *Conf Comput Vis Pattern Recognit)*. 2016: 2536–2544. https://doi.org/10.1109/CVPR.2016.278
20. Noroozi M, Favaro P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science; 9910. Springer; 2016. https://doi.org/10.1007/978-3-319-46466-4_5doi:10.48550/arXiv.1603.09246
21. Gidaris S, Komodakis N. Unsupervised representation learning by predicting image rotations. *Preprint*. Posted online on March 21, 2018. arXiv preprint arXiv:1803.07728. https://doi.org/10.48550/arXiv.1803.07728
22. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020:9729–9738.
23. Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM. Learning to compare: relation network for few-shot learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018:1199–1208.
24. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning, PMLR*; 2020;119:1597–1607.
25. Grill J-B, Strub F, Altché F, et al. Bootstrap your own latent - a new approach to self-supervised learning. *Advances in neural information processing systems*. 2020;33:21271–21284.
26. Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*. 2020;33:9912–9924.
27. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016:2424–2433. https://doi.org/10.1109/CVPR.2016.266
28. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. *Med Image Anal*. 2019;58:101539. https://doi.org/10.1016/j.media.2019.101539
29. Liu X, Zhang F, Hou Z, et al. Self-supervised learning: generative or contrastive. *IEEE Trans Knowl Data Eng*. 2023;35:857–876. https://doi.org/10.1109/TKDE.2021.3090866
30. Wang X, Yang S, Zhang J, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal*. 2022;81, 102559. https://doi.org/10.1016/j.media.2022.102559
31. Xie Y, Xu Z, Zhang J, Wang Z, Ji S. Self-supervised learning of graph neural networks: a unified review. *IEEE Trans Pattern Anal Mach Intell*. 2023;45: 2412–2429. https://doi.org/10.1109/TPAMI.2022.3170559
32. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018:7794–7803.
33. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: criss-cross attention for semantic segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019:603–612.
34. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H. Dual attention network for scene segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019:3146–3154.
35. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 7132–7141.
36. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS. Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017:5659–5667.
37. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: *International conference on machine learning*. PMLR; 2018: 2127–2136.
38. Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154–163. https://doi.org/10.1038/s41591-021-01620-2
39. Kather JN, Weis CA, Bianconi F, et al. Multi-class texture analysis in colorectal cancer histology. *Sci Rep*. 2016;6(1):27988. https://doi.org/10.1038/srep27988
40. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009. Accessed November 21, 2022. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf
41. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68–A77. https://doi.org/10.5114/wo.2014.47136
42. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. *Mach Learn Appl*. 2022;7:100198. https://doi.org/10.1016/j.mlwa.2021.100198
43. Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Med Image Anal*. 2021;70:102032. https://doi.org/10.1016/j.media.2021.102032
44. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat Med*. 2023;29(9):2307–2316. https://doi.org/10.1038/s41591-023-02504-3
45. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. *Nat Med*. 2024;30(3):863–874. https://doi.org/10.1038/s41591-024-02856-4
46. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w
47. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. *Pattern Recognit*. 2013;46(1): 243–256. https://doi.org/10.1016/j.patcog.2012.07.021
48. Ashari IF, Nugroho ED, Baraku R, et al. Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta. *J Appl Informat Comput*. 2023;7(1):95–103. https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/4947
49. Mughnyanti M, Efendi S, Zarlis M. Analysis of determining centroid clustering x-means algorithm with Davies-Bouldin index evaluation. *IOP Conf Ser: Mater Sci Eng*. 2020;725:012128. https://doi.org/10.1088/1757-899X/725/1/012128