Stain-Adaptive Self-Supervised Learning for Histopathology Image Analysis

Hai-Li Ye\textsuperscript{a,b,*}, Da-Han Wang\textsuperscript{a,b}

\textsuperscript{a}Department of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361000, CN
\textsuperscript{b}Fujian Provincial Key Laboratory of Pattern Recognition and Image Understanding, Xiamen 361000, CN

\textbf{A R T I C L E I N F O}

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013
Communicated by S. Sarkar

\textbf{2000 MSC:} 41A05, 41A10, 65D05, 65D17

\textbf{Keywords:} Computational Pathology, Self-Supervision, Deep Learning

\textbf{A B S T R A C T}

It is commonly recognized that color variations caused by differences in stains is a critical issue for histopathology image analysis. Existing methods adopt color matching, stain separation, stain transfer or the combination of them to alleviate the stain variation problem. In this paper, we propose a novel Stain-Adaptive Self-Supervised Learning (SASSL) method for histopathology image analysis. Our SASSL integrates a domain-adversarial training module into the SSL framework to learn distinctive features that are robust to both various transformations and stain variations. The proposed SASSL is regarded as a general method for domain-invariant feature extraction which can be flexibly combined with arbitrary downstream histopathology image analysis modules (e.g. nuclei/tissue segmentation) by fine-tuning the features for specific downstream tasks. We conducted experiments on publicly available pathological image analysis datasets including the PANDA, BreastPathQ, and CAMELYON16 datasets, achieving the state-of-the-art performance. Experimental results demonstrate that the proposed method can robustly improve the feature extraction ability of the model, and achieve stable performance improvement in downstream tasks.

© 2022 Elsevier B. V. All rights reserved.

1. Introducing

Pathological diagnosis\textsuperscript{[Epstein et al. (2006); Emad et al. (2008); Symmans et al. (2007)]} plays a critical role in clinical medicine because it provides objective evidence for the diagnosis, classification, and treatment of diseases, as well as for the judgment of disease progression, prognosis, and efficacy. Pathology is the process and principle of the occurrence and development of a disease, including the causes and regulations of the event of a condition, as well as the changes in the structure, function, and metabolism of cells, tissues, and organs during the process of an illness and their rules. In addition to providing diagnostic information, the phenotypic information contained in histology slides can be used for prognosis. Features such as nuclear atypia, degree of gland formation, presence of mitosis, and inflammation can all be indicative\textsuperscript{[Emad et al. (2008)]} of how aggressive a tumor is and may also allow predictions to be made about the likelihood of recurrence after surgery.

The analysis of whole-slide digital pathology images (WSIs)\textsuperscript{[Gurcan et al. (2009)]} is a challenging task because complex background with noises, the nuclei overlapping, variations in staining, etc. Among these problems, the problem of stain variation has been generally recognized as a critical issue that may deterioate the performance of the system\textsuperscript{[Rabinovich et al. (2003); Bejnordi et al. (2016); BenTaieb and Hamarneh (2018)]}, and has attracted much attention in recent years. The resolution of WSIs exceeds tens of thousands of pixels, so it is necessary to use a sliding window to divide them into patches during analysis. The pathological features of microscopes at different magnifications may be all-important for a task. Therefore information from multiple scales needs to be integrated. When the proportion of lesion area in WSIS is tiny, the lesion

\textsuperscript{*}Department of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361000, CN. E-mail addresses: yehl@mail.sustech.edu.cn (H. Ye).
information reflected by the pixels of the relevant area is minimal. All in all, the Analysis of WSIs is challenging in the following two respects. The first difficulty is that staining conditions vary greatly depending on the specimen and the hospital from which the sample was taken. Therefore, pathologists perform tumor region identification and subtype classification by carefully considering the different staining conditions. The last difficulty is that Pathological image data is relatively scarce. This is mainly due to the scarcity of some problematic cases and the high maintenance cost of WSIs. In addition, the current pathological image data sets are relatively independent, and data collection and labeling are carried out for different organs and lesions. Sample imbalance is more severe in pathological image analysis compared with other medical image analysis fields. However, some current work based on transfer learning and stain normalization successfully extract WSIs invariant features. But these methods are more complex and not applicable in the process of multiple transfers.

To address these common problems, we propose a simple and effective solution called Stain-Adaptive Self-Supervised Learning (SASSL) for Histopathology Image Analysis. Our goal is to offer a novel self-supervised method to achieve model adaptation to staining while extracting potential invariance features of WSIs. It reduces the impact of pathological image staining differences and can also use additional tasks to mine its monitoring information from large-scale pathological image data of different stains. The model aligns the distribution of all staining stains in common feature space to extract the invariant representation of the common staining stains. In addition, this module mines the supervision information of pathological image data itself, and the consistent features in pathological images can be learned like the self-supervision process. Its function is to extract the invariant features in the pathological image while aligning the staining stains of the pathological image to improve the performance of the downstream model of specific pathological image analysis. Because each downstream is considered to have a particular distribution of features, we do not directly replace the downstream feature extraction module as in the general self-supervision process. Instead, we want the generic feature extractor as an auxiliary branch. Its function is to extract the aligned invariant features from pathological images, and these features can further enrich the feature extraction ability of downstream models.

2. Related Work

In our review of related works, we first introduce three major canonical deep learning models based on the nature of tasks that are solved in digital histopathology: classification, regression, and segmentation-based models, and then we summarize the self-supervised learning based on contrastive learning and pathological image stain normalization. Finally, we will explain the difference and advancement of our work.

Pathological Image Classification: Image classification is a classic computer vision problem. Compared with the general visual image, the contour, and texture contained in the image block of WSIS are more complex. As the number of layers of the neural network increases, the problem of gradient vanishing and gradient explosion will be more obvious. In [He et al. (2016)], the authors propose a residual learning method, which uses the deep layer as the shallow layer Identity mapping to prevent the gradient from disappearing or exploding. In addition to similar structures [Huang et al. (2017); Tan and Le (2019); Ding et al. (2021)], others have a series of models based on attention mechanism include [Hu et al. (2018); Woo et al. (2018); Gao et al. (2021); Hou et al. (2021)]. However, the global classification of pathological images is more complicated than the scene image classification. The commonly used pathological image classification [Campanella et al. (2019); Gao et al. (2017); Tellez et al. (2018); Xu et al. (2019)] is based on multiple instances of a learning-based deep learning system that uses only the reported diagnoses as labels for training, thereby avoiding expensive and time-consuming pixel-wise manual annotations.

Pathological Image Regression: This kind of research focuses on the detection or localization of objects by directly regressing the likelihood of a pixel being the center of an object (e.g., cell or nucleus center). Scene image regression first assumes a linear relationship between the pixel values of two images in different periods. The regression and classification models are very similar in structure; the major difference lies at the end of the model. At present, the main research is in the target detection how to more accurately return to the area of the object (Xie and Tu (2017); Naylor et al. (2019a); Chen et al. (2016); Kashif et al. (2016);). Detection of cells or nuclei in histopathology images is challenging due to their highly irregular appearance and their tendency to occur as overlapping clumps, which results in difficulty in separating them as a single cell or a nucleus (Naylor et al. (2019b); Xie et al. (2015); Graham et al. (2019)). Therefore, end-to-end regression models can be used directly for tumor load assessment (Geert et al. (2018); Bandi et al. (2019); Ehteshami Bejnordi et al. (2017)) and lymphocyte assessment (Mohammad et al. (2017); Kurc et al. (2019); Lomachenko and Arandjelovic (2021)).

Pathological Image Segmentation: In recent years, the development of deep neural networks has achieved many breakthroughs in automatic histopathology image segmentation. FCN (Long et al. (2015)) promotes the use of end-to-end convolutional neural networks in semantic segmentation problems. U-Net (Ronneberger et al. (2015)) model constructs a complete set of encoder-decoder, which is widely used in medical image segmentation. And other follow-up work (Çiçek et al. (2016); Milletari et al. (2016); Zhou et al. (2018); Chen et al. (2018)). And also work on the features of the lesions, such as the liver cancer [Kim et al. (2021); Schmitz et al. (2021)]; Breast cancer [Liu et al. (2019); Kiani et al. (2020)]; lung cancer [Nicolas et al. (2018); Wang et al. (2020)]. In addition to segmentation based on global size, there are also segmentation based on nuclear level (Song et al. (2019); Hu et al. (2019)). Many researches also aim at optimizing loss to improve segmentation quality (Milletari et al. (2016); Kervade et al. (2021)). Most deep learning methods in digital pathology are applied on small-sized image patches rather than the entire WSI, restricting the model’s prediction ability to a narrow field of view.
Fig. 1. stain-adaptive self-supervised learning (SASSL) architecture. (Please note that the input image size does not need to be fixed value). In step 1 (blue area), samples from multiple pathological image data sets are input into SASSL for stain adversarial self-supervised learning. Compared with traditional self-supervised methods, we introduce a stains discrimination module (green area). For a detailed description of the SASSL process, please refer to Chapter 3.2. In step 2, the weights learned by SASSL will be transferred to downstream tasks to improve the feature extraction ability of the model. For detailed description, please refer to Chapter 3.3.

Self-Supervised Contrastive Learning: In recent years, self-supervised learning [Jing and Tian (2021); Liu et al. (2020); Chen and He (2021); Chen et al. (2020c)] as a new representational learning method has achieved many achievements. Self-supervised learning mainly uses auxiliary tasks to mine its supervised information from large-scale unsupervised data. Self-supervised trains the network with the constructed supervisory information so that the valuable representations for downstream tasks can be learned. After self-supervised training, the learned visual features can be further migrated to downstream tasks as pre-training models to improve performance and overcome over-fitting. We summarize them into three main categories according to their objectives: generative [Wang et al. (2014); Fabius et al. (2015); Xu et al. (2020)], contrastive [Chen et al. (2020a); Chen et al. (2020b); He et al. (2020); Chen et al. (2020c); Grill et al. (2020)], and adversarial [Radford et al. (2016); Zhang et al. (2016); Zhang et al. (2017)]. The generation-based or adversarial-based approach focuses more on the details of the pixel than on the more abstract underlying factors. Therefore, the method based on a contrastive pays more attention to the global semantic features of images and can learn more complex potential representations of images.

Pathological Image Stain Normalization: Stain normalization techniques have been an important preprocessing before most computer-aided diagnostic tasks, as slides from different institutions and even slides within the same institution but different batches may vary drastically in stain style. Many successful stain normalization techniques have been proposed in the discipline of computational pathology. The initial approach attempts to subtract each color channel’s mean and then divide by the standard deviation [Nishar et al. (2020)]. Conventional algorithms have demonstrated their effectiveness in the past, like a non-linear mapping approach that employs image-specific color deconvolution [Macenko et al. (2009)], or an estimation of the stain vectors using singular value decomposition (SVD) geodesic-based stain normalization technique [Reinhard et al. (2001)], and many others [Vahadane et al. (2016); Nadeem et al. (2020)]. However, due to the reliance on an expertly selected target image, the conventional methods are no more applicable to convolutional neural networks. More recently, generative adversarial networks have been widely adopted to reduce color variations between multiple data centers [Ke et al. (2021)]. Specifically, the StainGAN [Shaban et al. (2019)] have demonstrated their outstanding experimental results concerning stain separation and image information preservation.

In this paper, we propose a simple and effective solution called Stain-Adaptive Self-Supervised Learning (SASSL) for Histopathology Image Analysis. Our goal is to propose a novel self-supervised method to achieve model adaptation to staining while extracting potential invariance features of WSIs. The key contributions of our paper are:
The proposed SASSL method integrates a stains discriminator module into the SSL framework to perform adversarial learning. This method improves the adaptive ability of the model to WSIs staining changes and enables the model to learn to extract potential invariance features in WSIs.

Our SASSL method can be regarded as a general model for invariant feature extraction which can be flexibly combined with arbitrary downstream histopathology image analysis modules (e.g., Classification/Regression/Segmentation).

3. Methodology

Fig. 1 illustrates the overview of the proposed Stain-Adaptive Self-Supervised Learning (SASSL) framework for the pathological image analysis. It mainly consists of two steps: (1) In the first stage, we used multiple pathological image data sets for self-supervised learning without using downstream task supervised labels. We introduce a stains discriminative module so that the self-supervised module realizes the stains alignment of pathological images in the high-dimensional feature space and mines the potential standard features in pathological images. (2) We treat the encoder learned in the first stage as a universal feature extractor in the second stage. The generic feature extractor can extract general features in pathological images and parallel to the special feature extractor of the original downstream framework to extract and fuse image features. This process can be embedded into any downstream analysis framework of pathological images, which improves the analysis capability of the model robustly and has good mobility.

3.1. Self-Supervised Learning Framework

In order to realize stain-adaptive self-supervised learning of general data of pathological images, we first need to integrate pathological image data sets from different task scenes. We selected the three classic task scenes in the visual analysis of pathological images: classification, regression, and segmentation. The core of stains adaptation is to make the model unaffected by the stains changes of WSIs and achieve robust feature extraction for any stains situation. This idea is similar to the fine-grained identification scene, which contradicts small inter-class and large intra-class distances. We treat multiple pathological image datasets as a whole, including N WSI slices. We denote the multi-source dataset as \( \{(X_n, Y_n)\}_{n=1}^N \), where \( X_n \) is WSI and \( Y_n \) is downstream task label. The self-supervised learning in the first phase of our framework is an unsupervised process that does not need to use labels for downstream tasks.

When the downstream task is a classification, \( Y_n \) is the single hot coding of the classification label. Similarly, \( Y_n \) corresponds to linear predictive value and segmentation mask in regression and segmentation tasks, respectively. Since each WSI is too huge to feed into a CNN directly, a patch-based approach is usually employed. In addition, different downstream tasks have different requirements for the input patches size, so we hope the model can accept multi-scale patches. In this paper, we consider patches with 512 × 512 pixels. Therefore, each WSI can be regarded as a stain type, and the patches generated are the sample of this stain. Patches is denoted by \( P_n \) for \( n \in [N] \) and each patch \( p \in P_n \) belongs to the corresponding WSI. Stain-adaptive self-supervised learning aims to align the features of patches in different stains in high-dimensional space and mine the potential common features in pathological images.

The structural paradigm of SASSL is modeled on self-supervised learning based on contrastive learning. So, in the next section, we will briefly summarize the self-supervised contrast learning paradigm in preparation for the introduction of our stains antagonism module. Our approach is an extension of the underlying paradigm of comparative self-supervised learning that applies to approaches of the same paradigm type. Several SSL methods can be adopted as the base framework. Contrastive learning aims to “learn to compare” through a Noise Contrastive Estimation (NCE) objective formatted as:

\[
L = \mathbb{E}_{x,x^+,-}[\log\left(\frac{e^{f(x,x^*)}_{\theta}}{e^{f(x,x^*)}_{\theta} + e^{f(x,x^-)}_{\theta}}\right)],
\]

where \( x^+ \) is positive sample to \( x \), \( x^- \) is negative sample to \( x \) and \( f \) is an encoder (in this paper, it is a general feature extractor). The similarity measure and encoder may vary from task to task, but the framework remains the same. In general, with more negative sample pairs involved, we have the InfoNCE formulated as:

\[
L = \mathbb{E}_{x,x^+,-}[\log\left(\frac{e^{f(x,x^*)}_{\theta}}{e^{f(x,x^*)}_{\theta} + \sum_{i \neq m} e^{f(x,x_i)}_{\theta}}\right)],
\]

The deepening of related research mainly focuses on two process paradigms of contrastive learning. The first paradigm is mainly to increase the proportion of negative samples to strengthen comparative learning. We take MoCov2 (Chen et al. (2020c)) (Fig.2a) as an example. In MoCov2, researchers further develop the idea of leveraging instance discrimination via momentum contrast, substantially increasing the negative samples’ amount. The data goes through two or more different enhancements to get different views. These views have differences in shallow features but are consistent in deep representation. Where \( f_e(x) \) uses gradient descent training, \( f_{m_{\theta}}(x) \) updates momentum according to \( f_e(x) \), maintaining \( f_e(x) \) and \( f_{m_{\theta}}(x) \) similar but different state. After the view of the same sample, \( X \) is encoded by the encoder to get \( S_1 \) and \( S_2 \); the same sample vector features should be aligned. After each iteration, the previous sample is entered into the negative sample queue. Finally, there are positive sample pairs between \( S_1 \) and \( S_2 \); and negative sample pairs between \( S_1 \) and negative sample queue, using the Equation 2 InfoNCE loss to update the model parameters. The second paradigm discarded the form of negative sampling and achieved better results in self-supervised learning than InfoNCE (Gutmann and Hyvärinen (2012)). We take SimSiam (Chen and He (2021)) (Fig.2b) as an example. SimSiam abandons momentum update and changes to the twin network structure. An additional set of MLP is added to the \( f_e(x) \) encoder branch and turns off the gradient of another branch. Similarly, different views are passed through two encoders to get \( S_1 \) and \( S_2 \). SimSiam followed the regression paradigm to de-
sign the loss function:

\[ L = 1 - \frac{<S_1, S_2>}{||S_1||_2 \cdot ||S_2||_2}, \]  

(3)

### 3.2. Stains Discriminator Module

Both paradigms finally generate eigenvectors \( S_1 \) and \( S_2 \) of different views of samples. For pathologica images, the previous comparative learning process can only learn the potential characterization of patches but cannot realize the alignment of the staining stains between patches in high-dimensional space. For pathologica images, the previous comparative learning process can only learn the potential characterization of patches but cannot realize the alignment of the staining stains between patches in high-dimensional space. For pathological images, the previous comparative learning process can only learn the potential characterization of patches but cannot realize the alignment of the staining stains between patches in high-dimensional space.

**Stains Affinity Matrix.** We propose a stains affinity matrix for one mini-batch with \( N \) images to compute the affinity among different patch stains. The stains affinity matrix \( A : \mathbb{R}^{N,N} \) is obtained by multiplying \( l_1 : \mathbb{R}^{N,d} \) and \( l_2^T : \mathbb{R}^{d,N} \), where \( a_{i,j} \) represents the stains affinity of the \( i \) and \( j \) patch. For the stain discriminator, it is desired that, for the same image with a different augmentation view, the affinity between them is high. On the contrary, for different images, the affinity is low. When all patches in mini-batch are from different stains, the diagonal of the affinity matrix (the red block) is desired to be high while the others (the blue block) are desired to be low. In order to better illustrate this concept, we take the mini-batch \( N \) as four as an example, as shown in Fig.4. Only the blocks in the diagonal region are affinities when there are no patches of the same source. When the number of homologous patches increases, the affinity blocks increase. The relational matrix \( D \) is written as:

\[ R = (r_{i,j}) = \begin{cases} 
0, & p_i \neq p_j \\
1, & p_i = p_j 
\end{cases} \]

**Stain-Adaptive Training.** To reduce the model’s sensitivity to the stains, the adversarial training was carried out for the affinity matrix. On the contrary to the stains discriminator, for the generator, i.e., the feature learning of SSL, it is desired that all the images have high affinities despite what the augmentation/stains are. We treat the SSL part as generator \( G \) and the stains counter module as \( D \). Using the idea of GAN training, the Stain-Adaptive training object function will be:

\[ \min_{\theta_1, \theta_2}, \max_{\theta_3} L_{\text{adv}}(G, D) + L_{\text{ ssl}}(G), \]  

(4)
Further, we specifically write the loss expression of the antagonistic training process:

\[
L_G = L_{ssl} - \frac{||((1-R)A)||_1}{||(1-R)||_1} - \frac{||RA||_1}{||R||_1} = L_{ssl} - \frac{||A||_1}{N^2},
\]

(5)

\[
L_D = \frac{||((1-R)A)||_1}{||1-R||_1} - \frac{||RA||_1}{||R||_1},
\]

(6)

\( L_{ssl} \) in the above formula can be replaced by the self-supervised loss we introduced in Chapter 3.2. The above GAN training objective is regarded as a saddle point optimization problem, and gradient-based methods often accomplish the training. \( G \) and \( D \) have trained alternately from scratch so that they may evolve together. Where \( A \) is the affinity matrix of the model output, and \( R \) is the relationship matrix. \( R \) represents patches of red homology in Fig. 4, and \( 1 - R \) means patches of blue from different sources. For Eq. 5, the goal of the generator \( G \) is to ensure that the patch feature vectors of different colors are aligned simultaneously as self-learning. Therefore, the affinity of the feature vectors for any patch should be close (red and blue). Then all the values in the affinity matrix \( A \) should be as large as possible so that the discriminator \( G \) cannot discriminate the affinity of different slices to achieve the purpose of adaptation, that is, \( \frac{||A||_1}{N^2} \). For Eq. 6, the discriminator \( G \) only needs to be able to determine whether the patches are homologous. That is, the affinity of different source patches is suppressed, and the affinity of homologous patches is improved.

**Fig. 4. Overview of downstream tasks of pathological image analysis.**

### 3.3. Downstream Task Migration

The backbone after the above self-supervised training has a good representation extraction ability. Then, after the downstream tasks, the learned parameters can be migrated and fine-tuned, just like the general supervised learning process. Therefore, the ability of self-supervised learning is mainly reflected by the performance of downstream tasks. As shown in step 2 of Fig.1, we improve the general self-supervised migration process by introducing residual learning and attention mechanisms. The parameters of the generic feature extractor \( E_s \) after self-supervised learning are fixed, the gradient is turned off, and the features extracted by it are regarded as generic features. The parallel special feature extractor \( E_e \) obtains the special feature \( f_e \) by residual learning on the basis of the generic feature \( f_s \). Specifically, it can be expressed as the following process:

\[
f_{e_q}^l = E^l_s(f_{e_{q-1}}^l),
\]

(7)

\[
f_{e_q}^l = E^l_s(f_{e_{q-1}}^l) + \epsilon_{e_q}^l,
\]

(8)

Where, \( f_{e_q}^l \) is the general feature of the upper layer feature \( f_{e_{q-1}}^l \) through the \( l \) layer of \( E^l_s \). \( f_{e_q}^l \) is the special feature the upper layer feature \( f_{e_{q-1}}^l \) through the \( l \) layer of \( E^l_s \) and add the residual item \( \epsilon_{e_q}^l \). \( f_{e_q}^l \) of the same layer can be regarded as the feature modification of \( f_{e_q}^l \), that is, the special feature is the result of residual learning on the available feature. For the final general features \( f_e \) and special features \( f_e \), we combined them in the form of attentional mechanisms get \( f_e \):

\[
f_e = f_e + \epsilon_e + \sigma(f_e),
\]

(9)

Next, \( f_e \) can be easily embedded into the pathological image analysis downstream task introduced in related work Chapter 2. As shown in Fig.4, we briefly summarize this paper’s three forms of downstream tasks experiment. In the classification task, \( f_e \) inputs MLP to obtain each category’s confidence, and Softmax gets the classification result. In Fig.4, we illustrate the classification of ISUP Grade in breast cancer. Similarly, in the regression task, \( f_e \) also inputs MLP and finally inputs the nonlinear layer to output logistic regression results. In Fig.4, we illustrate the regression of the proportion of malignant and healthy cells. In the segmentation task, after \( f_e \) is input to the decoder, it is recovered and decoded layer by layer through deconvolution to generate the segmentation mask. In Fig.4, we exemplify tumor segmentation.

### 4. Experiment

#### 4.1. Datasets Preparation

Our experiments are mainly divided into two parts. First, we compared SASSL with other staining-normalized SOTA models. Secondly, we compared SASSL with other self-supervised methods for different pathological image downstream analysis tasks to thoroughly verify the adaptive ability and transfer robustness of SASSL to the staining changes of pathological slides. In the first part, We use a selection of histopathological datasets with 200 WSIs from 4 different tumors, 1) colon adenocarcinoma from TCGA-HCOAD(Kirk et al. (2016)), 2) rectum adenocarcinoma from TCGA-READ(Kirk et al. (2016)), 3) stomach adenocarcinoma from TCGA-STAD(Kirk et al. (2016)) and 4) breast cancer metastases from Camelyon16(Geert et al. (2018)), with a balanced distribution of 50 slides each. This part of WSIs were split into non-overlapped 224 × 224 patches at the magnitude of 20× to retain the high resolution. In the Second part, We evaluate the performance of our method on four pathological image analysis datasets: PANDA(Epstein et al. (2006)) data sets, BreastPathQ(Mohammad et al. (2017)) data sets, CAMELYON16(Geert et al. (2018)) data sets. Among the PANDA data sets is the classification data set, the BreastPathQ.
data set is the regression data set, and the CAMELYON16 data sets are segmentation data sets. This part of WSIs were split into non-overlapped 512 × 512 patches at the magnitude of 20x to retain the high resolution. All the datasets are publicly available online.

**TCGA-COAD (Kirk et al. 2016)** The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). The TCGA-COAD Data set consists of 983 WSIs of digitized H&E-stained. We divided 70% of the data into training sets and 30% into test sets. We split each WSI cut into patches at 20x magnification with a size of 224 × 224 with a balanced distribution of 50 slides each. We split into 49150 patches, the overall label for subordinate patches obtained by each WSIs.

**TCGA-READ (Kirk et al. 2016)** The Cancer Genome Atlas Rectum Adenocarcinoma (TCGA-READ) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). The TCGA-COAD Data set consists of 364 WSIs of digitized H&E-stained. We divided 70% of the data into training sets and 30% into test sets. We split each WSI cut into patches at 20x magnification with a size of 224 × 224 with a balanced distribution of 50 slides each. We split into 18200 patches, the overall label for subordinate patches obtained by each WSIs.

**TCGA-STAD (Kirk et al. 2016)** The Cancer Genome Atlas Stomach Adenocarcinoma (TCGA-STAD) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from The Cancer Genome Atlas (TCGA). The TCGA-COAD Data set consists of 758 WSIs of digitized H&E-stained. We divided 70% of the data into training sets and 30% into test sets. We split each WSI cut into patches at 20x magnification with a size of 224 × 224 with a balanced distribution of 50 slides each. We split into 37900 patches, the overall label for subordinate patches obtained by each WSIs.

**PANDA (Epstein et al. 2006)** Prostate cancer (PCa) is the second most common cancer in men in the world. Pathologists score the diagnosis based on the Gleason scoring system. In the Gleason scoring system, the rating is converted to an ISUP rating of 1-5. The ISUP grading plays a key role in determining how patients should be treated. The PANDA Data set consists of around 10,616 WSIs of digitized H&E-stained from two centers. There were 5,456 from Karolinska and 5,160 from radboud. We divided 80% of the data into training sets and 20% into test sets. We split each WSI cut into patches at 20x magnification with a size of 512 × 512. We split into 406,894 patches.

The overall label for subordinate patches obtained by each WSIs is the ISUP.

**BreastPathQ (Mohammad et al. 2017)** Breast cancer is the phenomenon of runaway proliferation of mammary epithelial cells under the action of various carcinogenic factors. Often referred to as the "pink killer," breast cancer is the number one malignancy in women. The assessment of cell structure is an important part of tumor load assessment. The number of cells in the tumor bed is defined as the percentage area of the entire tumor bed composed of tumor cells (invasive or in situ). In the BreastPathQ data set, 94 WSIs (69 training and 25 testing) that had been stained with H&E-stained were included, and 3700 ROI patches were extracted from the above WSI and marked with tumor cell fractions (0-100%) in these areas. The training set contains ROI patches 2579, the test set contains ROI patches 1121, and each patch is 512 × 512 in size.

**CAMELYON16 (Geert et al. 2018)** Automatic detection of lymph node metastasis has great potential to help pathologists and reduce their workload. The purpose of CAMELYON16 data set is to realize the automatic detection of cancer metastasis in lymph node images. CAMELYON16 include 400 hematoxylin-eosin (H&E) stained whole-slide images (WSIs) of lymph node sections. Among the 400 WSI sections, 270 WSIs (110 contain cancerous) were training sets and the remaining 130 WSIs (49 contain cancerous) were test sets. Even in cancer WSIs, normal tissues still account for a larger proportion. If too many normal WSIs are introduced for training, the model will not be able to learn the characteristics of cancer tissues well. Therefore, we only used WSIs containing cancer tissues for training and testing. For the first part of the experiment, we split each cancerous WSI cut into patches at 20x magnification with a size of 224 × 224. We split into 13500 patches in the train set, On the test set, we split into 6500 patches. For the second part of the experiment, we split each cancerous WSI cut into patches at 20x magnification with a size of 512 × 512. We split into 7853 patches in the train set, On the test set, we use the complete WSIs of the test set to evaluate the segmentation accuracy of the model.

In self-supervised training, We randomly selected 1000 WSIs from three data sets and filtered the background patches through OTSU. For each WSIs, we sampled 100 foreground patches in a balanced manner to ensure data balance between different stains. In the downstream task, the data set is segmented and trained according to the above description to verify the algorithm’s effectiveness. Therefore, we get a self-supervised training set containing 10^5 patches and use this training set for the self-supervised training. It is worth noting that in the self-monitoring training process, when the batch size is n, we balanced the number of patches belonging to the same slice, so as to avoid the situation that there are no homologous patches in the training process, that is, the relationship matrix is Fig3a.
Table 1. Comparison of the State-Of-The-Art of stain normalization benchmark in the four data test set. We use Acc(Accuracy) as a metric to evaluate classification performances with SASSL and SOTA methods. The last column is the average performance of the different models across the four data sets.

| Method | CAMELYON16 | TCGA = STAD | TCGA = COAD | TCGA = READ | Aggregated |
|--------|------------|-------------|-------------|-------------|------------|
| Reinhard et al. (2001) | 0.893 | 0.901 | 0.903 | 0.883 | 0.895 |
| Macenko et al. (2009) | 0.867 | 0.893 | 0.912 | 0.905 | 0.894 |
| Khan et al. (2014) | 0.912 | 0.902 | 0.893 | 0.911 | 0.904 |
| Vahadane et al. (2016) | 0.908 | 0.914 | 0.908 | 0.906 | 0.909 |
| Shaban et al. (2019) | 0.918 | 0.921 | 0.922 | 0.912 | 0.918 |
| Nishar et al. (2020) | 0.913 | 0.904 | 0.912 | 0.907 | 0.909 |
| Nadeem et al. (2020) [11] | 0.878 | 0.902 | 0.894 | 0.889 | 0.891 |
| Ke et al. (2021) | 0.923 | 0.927 | 0.921 | 0.925 | 0.924 |
| EfficientNet-B6 | 0.881 | 0.893 | 0.886 | 0.886 | 0.886 |
| EfficientNet-B6(SASSL) | 0.909 | 0.914 | 0.917 | 0.919 | 0.914 |
| EfficientNet-B1 | 0.911 | 0.923 | 0.915 | 0.917 | 0.916 |
| EfficientNet-B1(SASSL) | 0.942 | 0.952 | 0.953 | 0.957 | 0.951 |
| ResNet34 | 0.876 | 0.881 | 0.884 | 0.882 | 0.881 |
| ResNet34(SASSL) | 0.908 | 0.921 | 0.925 | 0.917 | 0.918 |
| ResNet50 | 0.864 | 0.886 | 0.874 | 0.879 | 0.876 |
| ResNet50(SASSL) | 0.926 | 0.944 | 0.935 | 0.932 | 0.934 |
| RepVGG16 | 0.894 | 0.903 | 0.902 | 0.908 | 0.901 |
| RepVGG16(SASSL) | 0.927 | 0.925 | 0.928 | 0.925 | 0.923 |
| RepVGG19 | 0.917 | 0.909 | 0.909 | 0.911 | 0.911 |
| RepVGG19(SASSL) | 0.926 | 0.931 | 0.928 | 0.931 | 0.929 |
| VIT | 0.852 | 0.854 | 0.88 | 0.855 | 0.858 |
| VIT(SASSL) | 0.887 | 0.89 | 0.881 | 0.875 | 0.883 |
| Swin | 0.855 | 0.868 | 0.867 | 0.867 | 0.864 |
| Swin(SASSL) | 0.893 | 0.905 | 0.893 | 0.887 | 0.894 |
| NASNet | 0.864 | 0.873 | 0.868 | 0.871 | 0.869 |
| NASNet(SASSL) | 0.897 | 0.901 | 0.899 | 0.895 | 0.898 |
| ProxylessNAS | 0.886 | 0.897 | 0.897 | 0.897 | 0.894 |
| ProxylessNAS(SASSL) | 0.904 | 0.922 | 0.912 | 0.923 | 0.915 |

4.2. Evaluation Metric

In the first part, We use Acc(Accuracy) as a metric to evaluate classification performances with SASSL and SOTA methods. In the second part, We use WQK(Weighted Quadratic Kappa), F1(F1 Score), and Acc(Accuracy) as a metric to evaluate classification performances. Weighted Quadratic Kappa measures the agreement between two outcomes. This metric typically varies from 0 (random agreement) to 1 (complete agreement). If there is less agreement than expected by chance, the metric may go below 0. The quadratic weighted kappa is calculated as:

$$kappa = 1 - \frac{\sum_i \sum_j w_{ij} O_{ij}}{\sum_i \sum_j w_{ij} E_{ij}}, \tag{10}$$

And then also, $F1(F1 \text{ Score})$ and Acc(Accuracy) was introduced as the evaluation criterion used to evaluate the classification result. We use $MSE$, $MAE$, and $R^2$ to verify regression performance. Finally, $PA$(Pixel Accuracy), $Dice$ coefficient, and $MIoU$(mean Intersection over Union) were used to evaluate the segmentation performance.

5. Results And Discussion

5.1. Comparison to state-of-the-art stain normalization frameworks

In the Frist part of experiment, We evaluated the proposed method with a couple of Sate-Of-The-Art(SOTA) stain normalization approaches proposed by Reinhard et al. [2001], Macenko et al. [2009], Khan et al. [2014], Vahadane et al. [2016], Shaban et al. [2019], Nishar et al. [2020], and Nadeem et al. [2020]. The demonstrated results are performed on test sets. We use the EfficientNet(Tan and Le [2019]), ResNet(He et al. [2016]), Swin(Swin Transformer) and ProxylessNAS(Swin Transformer).
Table 2. Comparison of the structural ablation of different self-supervised frameworks base on ResNet50 in the PANDA data set. In PANDA Dataset testing set got the following average scores. QWK represents Quadratic weighted kappa. Acc means Accuracy score. $F^3$ - $F^5$ represents the scores of different ISUP categories, and $F_{micro}$ means the micro F1 score of all categories.

| Method                  | $QWK_{it}$ | $QWK_{i}$ | $QWK_{r}$ | Acc  | $F^3$ | $F^4$ | $F^5$ | $F^2$ | $F^3$ | $F^4$ | $F^5$ | $F_{micro}$ |
|-------------------------|------------|------------|------------|------|-------|-------|-------|-------|-------|-------|-------|------------|
| BYOL                    | 0.878      | 0.846      | 0.865      | 0.598| 0.810 | 0.643 | 0.444 | 0.387 | 0.468 | 0.542 | 0.549 |            |
| BYOL(SASSL, w/o RL)     | 0.898      | 0.866      | 0.891      | 0.62 | 0.799 | 0.650 | 0.484 | 0.451 | 0.527 | 0.6   | 0.585 |            |
| BYOL(SASSL, w/ RL)      | 0.907      | 0.887      | 0.894      | 0.453| 0.860 | 0.668 | 0.467 | 0.444 | 0.538 | 0.640 | 0.603 |            |
| SimSiam                 | 0.872      | 0.856      | 0.856      | 0.578| 0.753 | 0.622 | 0.434 | 0.407 | 0.498 | 0.515 | 0.541 |            |
| SimSiam(SASSL, w/o RL)  | 0.903      | 0.891      | 0.894      | 0.641| 0.837 | 0.686 | 0.459 | 0.441 | 0.532 | 0.593 | 0.592 |            |
| SimSiam(SASSL, w/ RL)   | 0.910      | 0.903      | 0.896      | 0.645| 0.842 | 0.682 | 0.438 | 0.407 | 0.543 | 0.696 | 0.601 |            |
| SimCLR                  | 0.862      | 0.845      | 0.837      | 0.566| 0.748 | 0.607 | 0.456 | 0.416 | 0.425 | 0.493 | 0.524 |            |
| SimCLR(SASSL, w/o RL)   | 0.877      | 0.860      | 0.858      | 0.608| 0.829 | 0.647 | 0.443 | 0.425 | 0.441 | 0.553 | 0.556 |            |
| SimCLR(SASSL, w/ RL)    | 0.890      | 0.873      | 0.708      | 0.639| 0.808 | 0.709 | 0.480 | 0.474 | 0.487 | 0.573 | 0.588 |            |
| MoCo                    | 0.865      | 0.844      | 0.839      | 0.591| 0.782 | 0.634 | 0.478 | 0.442 | 0.432 | 0.5    | 0.545 |            |
| MoCo(SASSL, w/o RL)     | 0.883      | 0.865      | 0.865      | 0.610| 0.803 | 0.656 | 0.494 | 0.459 | 0.449 | 0.525 | 0.564 |            |
| MoCo(SASSL, w/ RL)      | 0.893      | 0.875      | 0.875      | 0.624| 0.806 | 0.691 | 0.498 | 0.439 | 0.457 | 0.589 | 0.580 |            |

Table 3. Comparison of the structural ablation of different self-supervised frameworks base on ResNet50 in the BreastPathQ test data set. MAE means Mean Absolute error, MSE means Mean Square Error, and $R^2$ means the R-Square.

| Method                  | MAE     | MSE     | $R^2$  |
|-------------------------|---------|---------|--------|
| BYOL                    | 0.198   | 0.060   | 0.425  |
| BYOL(SASSL, w/o RL)     | 0.190   | 0.058   | 0.441  |
| BYOL(SASSL, w/ RL)      | 0.186   | 0.055   | 0.474  |
| SimSiam                 | 0.187   | 0.056   | 0.468  |
| SimSiam(SASSL, w/o RL)  | 0.181   | 0.052   | 0.504  |
| SimSiam(SASSL, w/ RL)   | 0.179   | 0.051   | 0.512  |
| SimCLR                  | 0.195   | 0.060   | 0.431  |
| SimCLR(SASSL, w/o RL)   | 0.187   | 0.056   | 0.468  |
| SimCLR(SASSL, w/ RL)    | 0.180   | 0.051   | 0.508  |
| MoCo                    | 0.191   | 0.058   | 0.445  |
| MoCo(SASSL, w/o RL)     | 0.188   | 0.056   | 0.461  |
| MoCo(SASSL, w/ RL)      | 0.186   | 0.059   | 0.451  |

Table 4. Comparison of the structural ablation of different self-supervised frameworks base on ResNet50 in the CAMELYON16 test data set. PA means Pixel Accuracy, Dice means Dice coefficient and $MIoU$ means mean Intersection over Union.

| Method                  | PA     | Dice  | $MIoU$ |
|-------------------------|--------|-------|--------|
| BYOL                    | 0.842  | 0.816 | 0.724  |
| BYOL(SASSL, w/o RL)     | 0.874  | 0.823 | 0.734  |
| BYOL(SASSL, w/ RL)      | 0.893  | 0.835 | 0.758  |
| SimSiam                 | 0.905  | 0.840 | 0.766  |
| SimSiam(SASSL, w/o RL)  | 0.912  | 0.872 | 0.774  |
| SimSiam(SASSL, w/ RL)   | 0.929  | 0.906 | 0.839  |
| SimCLR                  | 0.886  | 0.842 | 0.728  |
| SimCLR(SASSL, w/o RL)   | 0.903  | 0.891 | 0.805  |
| SimCLR(SASSL, w/ RL)    | 0.912  | 0.902 | 0.782  |
| MoCo                    | 0.860  | 0.909 | 0.843  |
| MoCo(SASSL, w/o RL)     | 0.886  | 0.927 | 0.826  |
| MoCo(SASSL, w/ RL)      | 0.894  | 0.936 | 0.823  |
5.2. Performance of structural ablation on different self-supervised frameworks

Next, we verified the effects of our proposed SASSL method on different self-supervised frameworks. We selected SimCLR (Chen et al. (2020a)), MoCo (Chen et al. (2020c)), BYOL (Grill et al. (2020)), and SimSiam (Chen and He (2021)) as benchmarks. Among them, SimCLR (Chen et al. (2020a)) and MoCo2 (Chen et al. (2020c)) (Fig. 3a) belong to the self-supervised paradigm based on positive and negative samples, while BYOL (Grill et al. (2020)) and SimSiam (Chen and He (2021)) (Fig. 3b) belong to the self-supervised paradigm based on positive samples. We conducted comparative experiments on three data sets. The distribution of these three data sets belongs to different types of downstream tasks. The results of the PANDA data set are shown in Table 2, the results of the BreastPathQ data set are shown in Table 3, and the results of the CAMEONLY16 data set are shown in Table 4. In Table 1, we used ResNet50 as the backbone to perform the PANDA testing dataset score. In the method column, we compared the changes in the four self-supervised frameworks after the introduction of SASSL. QWK means Quadratic weighted kappa. ACC means Accuracy scores. $F_{1}^{5}$ represents the scores of different ISUP categories, and $F_{1}^{micro}$ means the micro $F_{1}$ score of all categories. It can be seen from the experimental results that SASSL can effectively improve downstream classification performance regardless of whether residual learning is performed. It can be seen from the QWK score transformation that the SASSL method not only improves the overall QWK score but also improves the classification accuracy of data from different central sources. This further shows that SASSL can significantly improve the model’s resistance to staining differences. For the four different self-supervised frameworks, QWK is improved by 1%-2% after using the SASSL method, and the overall improvement is between 2%-4% after the introduction of residual training. It is worth noting that in addition to robust improvement of samples from different data sources, SASSL can also generally improve the F1 value of each category.

In order to further analyze the improvement of the model feature learning ability by SASSL, we used CAM (Zhou et al. (2016)) to visualize the calorific value maps of different self-
supervised frameworks after using SASSL for comparison. As shown in Fig.5. In most current methods, the parameter initialization of the model is to use ImageNet’s pre-trained model weights. This can also bring a certain degree of improvement in the general scene data set. However, due to the significant difference between the characteristics of the medical image and the scene image, the advantage of this method in many medical image analysis works is not apparent. It can be seen in (Fig 5a) that the heating value effect obtained by initializing the weights of the ImageNet pre-training model is insignificant. The model pays more attention to the superficial features of pathological images, and there is no distinction between areas with apparent differences. Although the common self-supervised framework (Fig 5c,e.g. i) can effectively distinguish the critical areas in the pathological image, it has a higher calorific value corresponding to the tissue area. However, some areas with significant differences in dyeing still cannot produce higher heating values. After using SASSL (Fig 5d,f,h,j), common self-monitoring frameworks have a higher calorific response to key organizational areas. Among them, the improvement of the self-supervision method based on positive samples is the most obvious. It can be seen that the model has a high response to the tissue area and can effectively distinguish the unorganized area. After using SASSL for self-supervised training, downstream models can quickly locate key tissue regions and distinguish the difference between lesions and normal tissues.

In addition to classification tasks, SASSL is also effective in regression and segmentation tasks, as shown in Table 2. After adding SASSL for different self-supervised frameworks, the R2 of downstream models generally rises by 2%-4%. After the introduction of residual learning, it generally rises by 1%-7%, proving that SASSL is effective for downstream models. In the segmentation task, the result of CAMELYON16 is shown in Table 3. For different self-supervised frameworks, after adding SASSL, the R2 of downstream models generally rises by 1%-3%, and after the introduction of residual learning, it generally rises by 1%-6%. It can be seen from the results of use in Table 1-3 that SASSL is effective in different self-supervised methods. Furthermore, suitable for many types of tasks, SASSL can produce stable improvements.

5.3. Performance of structural ablation on different backbone

In previous experiments, we have demonstrated the effectiveness of SASSL on different self-supervised frameworks. In order to further prove the generalization of SASSL, we selected corresponding benchmark models for different task types, and we introduced SASSL for these benchmark models to analyze the changes in model performance. In order to make the experimental results more intuitive and avoid redundancy, the previous experimental results are integrated. Furthermore, the previous experimental results proved that the introduction of residual learning is also stable for the improvement of the model, so we selected the most stable performance of SimSiam as the self-supervised framework benchmark and performed residual learning. Under this benchmark, a unified comparison experiment was carried out. We all used SimSiam for self-supervised learning for the original SOTA model, which was compared with the addition of SASSL and residual learning.

| Method                      | $QWK_{all}$ | $QWK_s$ | $QWK_r$ | Acc  | $F_1^u$ | $F_1^i$ | $F_1^t$ | $F_1^p$ |
|-----------------------------|-------------|---------|---------|------|---------|---------|---------|---------|
| EfficientNet-B0             | 0.868       | 0.832   | 0.854   | 0.603| 0.799   | 0.677   | 0.414   | 0.412   | 0.419   | 0.522   | 0.540   |
| EfficientNet-B0(SASSL)      | 0.898       | 0.866   | 0.891   | 0.62 | 0.799   | 0.650   | 0.484   | 0.451   | 0.527   | 0.6    | 0.585   |
| EfficientNet-B1             | 0.935       | 0.917   | 0.923   | 0.701| 0.885   | 0.761   | 0.526   | 0.505   | 0.534   | 0.681   | 0.649   |
| ResNet34                    | 0.832       | 0.795   | 0.801   | 0.496| 0.640   | 0.543   | 0.378   | 0.340   | 0.330   | 0.594   | 0.471   |
| ResNet34(SASSL)             | 0.871       | 0.846   | 0.850   | 0.600| 0.772   | 0.66    | 0.457   | 0.407   | 0.440   | 0.456   | 0.550   |
| RestNet50(SASSL)            | 0.910       | 0.903   | 0.896   | 0.645| 0.842   | 0.682   | 0.438   | 0.407   | 0.543   | 0.696   | 0.601   |
| RepVGG16                    | 0.743       | 0.708   | 0.676   | 0.426| 0.447   | 0.485   | 0.375   | 0.373   | 0.430   | 0.330   | 0.407   |
| RepVGG16(SASSL)             | 0.756       | 0.719   | 0.735   | 0.578| 0.753   | 0.622   | 0.454   | 0.407   | 0.498   | 0.515   | 0.541   |
| RepVGG19                    | 0.773       | 0.702   | 0.753   | 0.432| 0.374   | 0.551   | 0.357   | 0.352   | 0.402   | 0.342   | 0.396   |
| RepVGG19(SASSL)             | 0.768       | 0.732   | 0.700   | 0.460| 0.582   | 0.522   | 0.381   | 0.360   | 0.350   | 0.380   | 0.429   |
| VIT                         | 0.755       | 0.735   | 0.671   | 0.477| 0.394   | 0.599   | 0.451   | 0.517   | 0.277   | 0.204   | 0.418   |
| VIT(SASSL)                  | 0.799       | 0.709   | 0.715   | 0.475| 0.414   | 0.568   | 0.449   | 0.493   | 0.453   | 0.203   | 0.431   |
| Swin                        | 0.793       | 0.772   | 0.743   | 0.573| 0.465   | 0.709   | 0.529   | 0.602   | 0.476   | 0.537   | 0.525   |
| Swin(SASSL)                 | 0.833       | 0.806   | 0.779   | 0.607| 0.682   | 0.716   | 0.493   | 0.586   | 0.506   | 0.646   | 0.555   |
| NASNet                      | 0.830       | 0.768   | 0.825   | 0.554| 0.700   | 0.586   | 0.462   | 0.398   | 0.454   | 0.561   | 0.527   |
| NASNet(SASSL)               | 0.861       | 0.813   | 0.851   | 0.527| 0.768   | 0.508   | 0.384   | 0.308   | 0.387   | 0.578   | 0.489   |
| ProxylessNAS                | 0.855       | 0.815   | 0.844   | 0.593| 0.710   | 0.650   | 0.499   | 0.436   | 0.485   | 0.584   | 0.561   |
| ProxylessNAS(SASSL)         | 0.882       | 0.853   | 0.869   | 0.578| 0.756   | 0.599   | 0.458   | 0.410   | 0.473   | 0.606   | 0.550   |

Table 5. Comparison of the structural ablation of different classifier baseline in the PANDA data set. In PANDA Dataset Testing set got the following average scores. $QWK$ means Quadratic weighted kappa. Acc means Accuracy score. $F_1^u-F_1^p$ represents the scores of different ISUP categories, and $F_1^{micro}$ means the micro F1 score of all categories.

---

Given-name Surname et al. Medical Image Analysis (2022) 11
Table 6. Comparison of the structural ablation of different different regressor baseline in the BreastPathQ data set.

| Method                | MAE  | MSE  | $R^2$ |
|-----------------------|------|------|-------|
| EfficientNet-B0        | 0.198| 0.060| 0.425 |
| EfficientNet-B0(SASSL)| 0.190| 0.058| 0.441 |
| EfficientNet-B1        | 0.187| 0.056| 0.468 |
| EfficientNet-B1(SASSL)| 0.181| 0.052| 0.504 |
| ResNet34               | 0.191| 0.058| 0.448 |
| ResNet34(SASSL)        | 0.183| 0.053| 0.491 |
| ResNet50               | 0.187| 0.056| 0.468 |
| ResNet50(SASSL)        | 0.179| 0.051| 0.512 |
| RepVGG16               | 0.225| 0.077| 0.425 |
| RepVGG16(SASSL)        | 0.204| 0.066| 0.371 |
| RepVGG19               | 0.205| 0.067| 0.363 |
| RepVGG19(SASSL)        | 0.187| 0.058| 0.441 |
| VIT                    | 0.199| 0.063| 0.399 |
| VIT(SASSL)             | 0.181| 0.052| 0.504 |
| Swin                   | 0.235| 0.083| 0.212 |
| Swin(SASSL)            | 0.209| 0.067| 0.357 |
| NASNet                 | 0.195| 0.060| 0.431 |
| NASNet(SASSL)          | 0.187| 0.056| 0.468 |
| ProxylessNAS           | 0.191| 0.058| 0.445 |
| ProxylessNAS(SASSL)    | 0.188| 0.056| 0.461 |

PANDA data set is much larger than that of the BreastPathQ data set. This allows the classification model to learn features well even when the initial weights are not good. In the regression task, the model lacks enough samples for feature extraction due to the small number of data set samples. Therefore, the initial weight is significant when the number of samples is small, which highlights the improvement of the model performance by SASSL. In order to better illustrate the feature learning of the regression task, we visualized the initial heating value changes of different SOTA models after SASSL was introduced, as shown in Fig.6. After SASSL is added, the difference in calorific value between the different tissue areas of the model is more obvious, which can help the model locate the lesion and key areas more quickly.

Regression tasks. Since the downstream model structure of the classification task and the regression task is very similar, we still use EfficientNet(Tan and Le (2019)), ResNet(He et al. (2016)), RepVGG(Ding et al. (2021)), VIT(Dosovitskiy et al. (2021)), Swin(Liu et al. (2021)), NASNet(Zoph et al. (2018)), and ProxylessNAS(Cai et al. (2019)) model as the SOTA benchmark, and sample the same comparison form as the classification task. Compare the original SimSiam with the results of adding SASSL and residual learning. The experimental results are shown in Table 6. From the results, we can see that, similar to the classification task, each SOTA model has improved steadily after the introduction of SASSL. Among them, the improvement of VIT(Dosovitskiy et al. (2021)) and Swin(Liu et al. (2021)) is still very significant. $R^2$ increased by 1% and 14%, respectively. Other benchmark models were similar to the classification task, with $R^2$ improvements of 2%-8%, respectively. One of the main reasons why the regression task is more effective than the classification task is that the sample size of the
are relatively novel segmentation models based on the Transformer structure. PRANet is the current SOTA model for segmenting polyps and foot ulcers. It has achieved good results with efficient attention feature modules. The experimental results are shown in Table 7. It can be seen that the performance of the above segmentation model has been significantly improved after adding SASSL. For SETR (Zheng et al. (2021)) and TransUNet (Chen et al. (2021)) based on the Transformer structure, the performance of the model is significantly improved after SASSL is added, and the Dice coefficient is increased by 7% and 5%, respectively. The Dice coefficient of DeepLabV3Plus has also increased a lot, an increase of 6.6%. The Dice coefficient of other models also increased by 1%-4%. In order to better analyze the improvement direction of the model for the segmentation structure, we visualized the segmentation results of each model for the patch, as shown in Fig. 7. SASSL can significantly improve the discriminative of the segmentation model for different regions in the image, and this kind of improvement is robust. It will not be subject to large fluctuations caused by color differences. The visualized results of the overall segmentation results of WSIs are shown in Fig. 8. From the perspective of the overall segmentation effect, after the introduction of SASSL, the model has significantly improved the segmentation effect of difficult-to-segment regions, reducing the probability of many such regions being segmented as false positives. Furthermore, the edges of the segmentation result are smoother, effectively reducing the appearance of jagged edges.

### 5.4. Conclusion

Computer-aided diagnosis of histopathological images can improve accuracy while reducing the burden on pathologists. This research proposes a stain-adaptive self-supervised learning framework (SASSL) method. Under un-supervised conditions, stain-adaptive self-supervised learning enables the model to adapt to staining while extracting potential invariance features of WSIs. SASSL automatically mines the potential features of pathological image data from multiple data sets, improves the model’s ability to extract deep features, and proves the effectiveness of SASSL on various types of downstream tasks. More importantly, SASSL training methods and pre-training weights can be extended to other research work for histopathological image analysis.

Although the performance of each model has been steadily improved after the introduction of SASSL, the model with the Transformer structure still has the potential for further improvement. By referring to Transformer’s self-supervised learning work in the NLP direction, the model based on Transformer structure can be generalized in pathological image analysis problems.

### Acknowledgments

Acknowledgments should be inserted at the end of the paper, before the references, not as a footnote to the title. Use the unnumbered Acknowledgements Head style for the Acknowledgments heading.

---

**Table 7.** Comparison of the structural ablation of different segmentation baseline in the CAMELYON16 data set.

| Method                  | PA    | Dice  | MIoU  |
|-------------------------|-------|-------|-------|
| UNet++                  | 0.905 | 0.840 | 0.766 |
| UNet++(SASSL)           | 0.929 | 0.906 | 0.839 |
| DeepLabV3Plus           | 0.876 | 0.883 | 0.791 |
| DeepLabV3Plus(SASSL)    | 0.905 | 0.923 | 0.857 |
| PRANet                  | 0.935 | 0.933 | 0.867 |
| PRANet(SASSL)           | **0.948** | **0.946** | **0.898** |
| SETR                    | 0.798 | 0.829 | 0.784 |
| SETR(SASSL)             | 0.850 | 0.894 | 0.809 |
| TransUNet               | 0.813 | 0.843 | 0.808 |
| TransUNet(SASSL)        | **0.883** | **0.899** | **0.818** |
bidirectional encoder representations from transformers (bert)-based ranking for concept normalization. J. Am. Medical Informatics Assoc. 27(10), 1510–1519. doi:10.1093/jamia/ocaa080

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), European Conference Computer Vision (ECCV), pp. 649–666. doi:10.1007/978-3-319-46487-9_40

Zhang, R., Isola, P., Efros, A.A., 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction, in: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 645–654. doi:10.1109/CVPR.2017.76

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 6881–6890. doi:10.48550/arXiv.2012.15840

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929. doi:10.1109/CVPR.2016.319

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention (MICCAI) DLMIA workshop, pp. 3–11. doi:10.1007/978-3-030-00889-5_1

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8697–8710. doi:10.1109/CVPR.2018.00907