bioengineering

MDPI

Article

RCKD: Response-Based Cross-Task Knowledge Distillation for Pathological Image Analysis

Hyunil Kim [1], Tae-Yeong Kwak [1], Hyeyoon Chang [1], Sun Woo Kim [1] and Injung Kim [2,*]

[1] Deep Bio Inc., Seoul 08380, Republic of Korea; hikim@deepbio.co.kr (H.K.); tykwak@deepbio.co.kr (T.-Y.K.); hychang@deepbio.co.kr (H.C.); swkim@deepbio.co.kr (S.W.K.)
[2] School of Computer Science and Electrical Engineering, Handong Global University, Pohang 37554, Republic of Korea
* Correspondence: ijkim@handong.edu

Abstract: We propose a novel transfer learning framework for pathological image analysis, the Response-based Cross-task Knowledge Distillation (RCKD), which improves the performance of the model by pretraining it on a large unlabeled dataset guided by a high-performance teacher model. RCKD first pretrains a student model to predict the nuclei segmentation results of the teacher model for unlabeled pathological images, and then fine-tunes the pretrained model for the downstream tasks, such as organ cancer sub-type classification and cancer region segmentation, using relatively small target datasets. Unlike conventional knowledge distillation, RCKD does not require that the target tasks of the teacher and student models be the same. Moreover, unlike conventional transfer learning, RCKD can transfer knowledge between models with different architectures. In addition, we propose a lightweight architecture, the Convolutional neural network with Spatial Attention by Transformers (CSAT), for processing high-resolution pathological images with limited memory and computation. CSAT exhibited a top-1 accuracy of 78.6% on ImageNet with only 3M parameters and 1.08 G multiply-accumulate (MAC) operations. When pretrained by RCKD, CSAT exhibited average classification and segmentation accuracies of 94.2% and 0.673 mIoU on six pathological image datasets, which is 4% and 0.043 mIoU higher than EfficientNet-B0, and 7.4% and 0.006 mIoU higher than ConvNextV2-Atto pretrained on ImageNet, respectively.

Keywords: deep learning; nuclei segmentation; knowledge distillation; contrastive learning; self supervised learning

check for updates

Citation: Kim, H.; Kwak, T.-Y.; Chang, H.; Kim, S.W.; Kim, I. RCKD: Response-Based Cross-Task Knowledge Distillation for Pathological Image Analysis. *Bioengineering* **2023**, *10*, 1279. https://doi.org/10.3390/bioengineering10111279

Academic Editors: Teng Grace Zhang, Jason Pui Yin Cheung and Tianjiao Zeng

Received: 11 September 2023
Revised: 19 October 2023
Accepted: 29 October 2023
Published: 2 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background

Pathological image analysis aims to extract useful information from pathological images commonly acquired through a whole slide scanner or camera. It covers various tasks, such as classification, segmentation, and detection of cells, nuclei, or cancerous regions, and is one of the core technologies for computer-aided diagnosis. Since deep learning exhibited outstanding performance in the ImageNet challenge [1], researchers have actively applied deep learning to pathological image analysis. Deep learning showed excellent performance in multiple challenges such as mitosis detection [2,3], breast cancer classification [4], and gland segmentation [5]. Currently, deep learning is widely used as a core algorithm in many pathological image analysis challenges [6]. In spite of that, there is a lot of room for improvement.

One of the main challenges is the difficulty of building a large-scale dataset for each pathological image analysis task. Collecting a large amount of data is restricted by privacy concerns. Moreover, labeling pathological images is more difficult and expensive than ordinary images. For example, training a deep learning model that predicts the aggressiveness of a tumor requires a dataset containing mitosis counting, cell segmentation labels, and the

patient's prognosis. Building such a dataset consumes significant time and effort from skilled pathologists [7].

Transfer learning is widely used in pathological image analysis to overcome the scarcity of labeled data and achieve high performance. A widely used approach is the pretraining and fine-tuning strategy, which first learns general knowledge by pretraining a model for an upstream task that can share knowledge with the target task using a large dataset, and then fine-tunes the pretrained model for the target task, also called the downstream task, using a relatively small target dataset [8,9]. However, because the characteristics of pathological images differ significantly from those of general images, transferring knowledge learned from a general image dataset, such as ImageNet [10], to a pathological image analysis task is less effective than ordinary transfer learning settings.

As an alternative, researchers have actively studied to learn features from unlabeled pathological images [11–17]. Recently, various unsupervised and self-supervised learning techniques, such as contrastive learning and masked autoencoders, have exhibited excellent performance in many computer vision tasks [18]. However, self-supervised learning algorithms do not perform as well in pathological image analysis as it does in other areas of computer vision, as described in Section 1.2.1. As a result, none of the existing supervised or self-supervised pretraining methods showed sufficient performance in pathological image analysis.

In addition, pathological images generally have significantly higher resolution than ordinary images, resulting in substantial increases in computational and memory requirements. For example, The Cancer Genome Atlas (TCGA) dataset [19] consists of images with a resolution of 20,000 × 40,000 pixels, which is hundreds of times the size of images in conventional datasets. Most pathological image analysis models decompose the whole-slide images (WSI) into patches to reduce the overhead. Nevertheless, memory and computational load is still an important issue in pathological image analysis. Many studies have been conducted to reduce deep learning models, but more research is needed to process high-resolution pathological images. Consequently, to achieve high performance in pathological image analysis in a general computing environment, we need not only a pretraining method that is effective in learning knowledge from unlabeled pathological images but also a lightweight architecture to reduce computational and memory overhead.

1.2. Related Work

In this subsection, we briefly introduce prior studies on self-supervised learning for pathological image analysis and efficient network architectures. We also present prior work on knowledge distillation and visual attention models on which the proposed methods are based.

1.2.1. Self-Supervised Learning for Pathological Image Analysis

Inspired by the success of self-supervised learning (SSL) in computer vision, many researchers have applied these techniques to pathological image analysis. Boyd et al. [11] applied a generative model-based learning method called visual field expansion. Ciga et al. [12] and Dehaene et al. [13] applied contrastive learning techniques such as SimCLR [20] and MoCoV2 [21] to pathological image analysis.

However, existing self-supervised pretraining techniques are less effective in pathological image analysis than in other computer vision fields. Zhang et al. [14] and Li et al. [15] reported analysis results suggesting that contrastive learning techniques are less effective for pathological image analysis. Koohbanani et al. [16] show a significant difference in performance between domain-agnostic and domain-specific tasks, suggesting that pathological images should be analyzed using pathological image-specific learning methods. Lin et al. [17] suggest a method to improve the performance of contrastive learning on pathological images by increasing the self-invariance, intra-invariance within a WSI, and inter-invariance across WSIs of the feature. However, their method requires clustering the feature vectors in each epoch, resulting in a significant increase in memory requirements.

1.2.2. Knowledge Distillation

Knowledge distillation (KD) is a technique for improving the training of a relatively small student model by exploiting the knowledge of a large and powerful teacher model. KD has been widely used to compress heavy models or to improve the performance of lightweight models. In the early days of KD, the student model learned to mimic the logits of the teacher model for each training sample [22], which is called response-based knowledge distillation (RKD). In a subsequent study, Adriana et al. [23] proposed feature-based knowledge distillation, which trains the student model using the intermediate features of the teacher model, enabling a more accurate approximation of the teacher model [24]. However, KD is primarily used under the assumption that the teacher and student models perform the same task in the same domain. KD performance often severely decreases when the task or domain of the student model differs from that of the teacher model. Li et al. [25] argued that the limitation comes from the fact that KD mainly transfers knowledge about global representation, and KD is less effective in transferring knowledge about local representation. To address this problem, they suggested using sub-modules to complement local knowledge.

A few previous studies apply KD to pathological image analysis. DiPalma et al. [26] propose a method for improving the computational efficiency of applying KD between models with the same structure but different input resolutions, and Javed et al. [27] proposed using additional modules to transfer knowledge stage-by-stage to learn robust tissue heterogeneity. Zhang et al. [28] presented a method of distilling knowledge from teacher models trained on diverse pathological datasets to help the student model learn the characteristic features of various pathological images.

However, these studies aim to reduce the size of a high-performance model that already exists and do not improve the performance of the high-performance model. It is hard to improve the performance of a model for a pathological image analysis task through conventional KD. Since the student model distills knowledge from the teacher model, a strong teacher model is a key requirement for KD. In pathological image analysis, where building a large-scale dataset is difficult, it is challenging to build a teacher model that performs the same task as the student model and is powerful enough to guide the learning of the student model.

1.2.3. Efficient Network Architectures for Image Analysis

There is a large body of previous work on the design of efficient network architectures. VGG16 [29] demonstrated that stacking multiple convolution filters with a kernel size of 3×3 can approximate a large kernel in image classification tasks. SqueezeNet [30] reduced the number of parameters using 1×1 convolutions and squeeze-expand modules, achieving similar performance to AlexNet with $50\times$ fewer parameters. ResNet [31] proposed a bottleneck structure to reduce the number of parameters in convolution blocks. MobileNet [32] reduced the number of parameters by up to 11% by decomposing convolution operations into a depthwise convolution and a pointwise convolution.

MobileNetV2 [33] achieved outstanding performance with a small number of parameters and computation using an inverted residual block and a linear bottleneck. Most of the recent architectures based on convolution or self-attention, such as EfficientNet [34], EfficientNetV2 [35], CoAtNet [36], ConvNext [37], ConvNextV2 [38], EfficientFormer [39], and EfficientFormerV2 [40], that exhibited good performance on ImageNet adopt the MBConv module proposed in [33]. In particular, CoAtNet proposes an architecture that combines CNN and self-attention. ConvNext achieved higher performance than Swin-B [41] of similar size by modernizing ResNet with several recent techniques such as the AdamW optimizer, a patchfy stem, large kernels of 7×7 size, fewer activation functions, and layer normalization.

1.2.4. Visual Attention

The attention mechanism allows the model to concentrate on essential features by assigning a weight to each feature based on its relevance. In general image processing, attention mechanisms are often categorized into channel attention, which learns the feature type to focus on, and spatial attention, which learns the locations of important features. Different channels of a feature map in a deep neural network represent different objects or concepts [42]. Hu et al. [43] proposed SENet, which estimates the importance of each channel and scales the channels accordingly. Gao et al. [44] pointed out that SENet is a simple structure designed to focus on important global information, so a more sophisticated structure is needed to focus on details better. Lee et al. [45] further reduced the size of SENet by applying a lightweight channel-wise fully-connected (CFC) layer. However, channel attention models can only learn what to focus on, but not where. Chen et al. [42], Park et al. [46], Woo et al. [47] demonstrated that using a combination of spatial attention and channel attention is superior to using channel attention alone. Wang et al. [48] showed that utilizing self-attention-based spatial attention can improve the performance of CNNs.

Dosovitskiy et al. [49] propose a vision Transformer (ViT), which modifies the Transformer network to fit image processing and achieves higher performance than CNN on the ImageNet dataset for the first time. Subsequent studies propose various image processing models based on self-attention, such as SwinTransformer [41], which extends ViT to a multi-scale structure, and CoAtNet, which combines separable convolution and multi-head self-attention (MSA).

1.3. Research Objective

In this study, we aim to develop a pretraining method and a lightweight network architecture to overcome the aforementioned challenges and improve the performance and efficiency of various pathological image analysis tasks. To this end, we propose the Response-based Cross-task Knowledge Distillation (RCKD) framework to learn knowledge from unlabeled pathological images using a high-performance model developed for a different task. We also propose the Convolutional neural network with Spatial Attention by Transformer (CSAT), an effective and efficient architecture designed as a backbone network for high-resolution image analysis. CSAT integrates multiple techniques that have shown to be effective in recent studies on lightweight architectures and further improves the performance through a novel Spatial Attention by Transformer (SAT) module. We expect that the results of this study will help reduce the cost of pathological image analysis and improve diagnostic performance by providing pathologists and researchers with fast and efficient diagnostic and research methods.

2. Materials and Methods

2.1. Datasets

In this study, we used different datasets for three specific purposes. For the pretraining of the model, we used the TCGA dataset, which consists of a large number of high-resolution pathological images. For fine-tuning and evaluation of the downstream tasks, we used six datasets described in Section 2.1.2. In addition, we used the ImageNet dataset to evaluate the performance and efficiency of network architectures in the analysis of general images not limited to pathological images.

2.1.1. Pretraining Dataset

TCGA dataset is one of the largest publicly available cancer genome datasets collected primarily for use in the diagnosis, treatment, and prevention of cancer. TCGA dataset includes more than 20,000 WSIs of stained tissue samples that belong to 33 different cancer types. TCGA dataset is widely used in various fields of pathological image analysis. From the TCGA dataset, we collected 11,716 WSIs from 32 different types of cancer. We excluded the formalin-fixed paraffin-embedded (FFPE) slide images because their quality

was poor. These samples were collected from a variety of organs, including the breast, brain, ovary, lung, kidney, prostate, stomach, and liver. Then, we cut the selected WSIs into non-overlapping patches of 1024 × 1024 size at 20× magnification. We removed patches whose average intensity values are not in the range of $[50, 245]$ because most of such patches consist of backgrounds rather than tissues. In this way, we collected 9,229,324 tissue patches. The types of studies and the number of whole slide imaging (WSI) are presented in Table A1. Among them, we randomly selected 400,000 patches and used them for pretraining.

2.1.2. Downstream Tasks and Datasets

For the fine-tuning and evaluation of the pretrained models, we used four segmentation datasets and two segmentation datasets. For the classification tasks, we used BACH (microscopy) [50], CRC [51], BreakHis [52], and Lymph [53] datasets.

- **The breast cancer histology images (BACH) microscopy dataset** contains 400 hematoxylin and eosin (H&E) stained microscopy image patches categorized into four classes: normal, benign, in situ carcinoma, and invasive carcinoma. Each class has 100 training image patches. The average patch size is 2048 × 1536 pixels.
- **The colorectal cancer (CRC) dataset** was collected for the classification of tissue areas from H&E stained WSIs of colorectal adenocarcinoma patients. It includes a total of 100,000 non-overlapping image patches in the training dataset extracted from 86 WSIs of cancer and normal tissues. The average patch size is 224 × 224 pixels. The patches were manually labeled by pathologists into nine tissue classes: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). The validation dataset includes 7180 image patches extracted from 50 colorectal adenocarcinoma patients.
- **The breast cancer histopathological (BreakHis) dataset** was collected for binary breast cancer classification. It contains a total of 7909 breast tumor tissue image patches from 82 patients, consisting of 2480 benign and 5429 malignant tumor patches. BreakHis includes four types of benign tumors: adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma, as well as four types of malignant tumors: ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. The average patch size is 700 × 460 pixels.
- **The Lymph dataset** was collected for the classification of malignant lymph node cancer. It provides a total of 374 training image patches, with 113 image patches of chronic lymphocytic leukemia (CLL), 139 image patches of follicular lymphoma (FL), and 122 image patches of mantle cell lymphoma (MCL). The average patch size is 1388 × 1040 pixels.

For the segmentation tasks, we used the BACH (WSI) [50] and GlaS [54] datasets.

- **The BACH WSI dataset** consists of 10 WSIs with an average size of $40,517 \times 58,509$ scanned at a resolution of 0.467 µ/pixel. It includes pixel-wise annotations in four region categories (normal, benign, in situ carcinoma, and invasive carcinoma). In this study, we cut the WSIs into 4,453 non-overlapping patches of 1024 × 1024 size at 10× magnification.
- **The gland segmentation in colon histology images (GlaS) dataset** is the benchmark dataset for the Gland Segmentation Challenge Contest at MICCAI in 2015. It consists of 165 image patches derived from 16 H&E stained histological sections of stage T3 or T42 colorectal adenocarcinoma. Each section originates from a different patient. The WSIs were digitized at a pixel resolution of 0.465 µm. The GlaS dataset consists of 74 benign and 91 malignant gland image patches, each with pixel-wise annotations. The average patch size is 775 × 522 pixels.

Since the BACH (microscopy), BreakHis, and Lymph datasets only release training sets to the public, we randomly split their training sets by 6:2:2 and used them for training, validation, and testing, respectively. The CRC and GlaS datasets provide the training

and validation sets but not the test set. We used 80% of their training sets for training and 20% for validation and measured the performance on the validation set. The BACH (WSI) dataset consists of ten WSIs. We used five of them for training, three for validation, and two for testing. Table 1 summarizes the task, classes, data size, number of patches, magnification ratio, and size of patches for the downstream datasets.

Table 1. Pathological image datasets for target tasks. (BACH: the BreAst Cancer Histology dataset, CRC: ColoRectal Cancer dataset, BreakHis: the Breast cancer Histopathological dataset, GlaS: the Gland Segmentation in colon histology images dataset).

Datasets	Tasks	Classes	Data Size	Number of Patches [Train, Validation, Test]	Magnification/ Patch Size
BACH [50] (microscopy)	Breast cancer subtype classification	Normal, Benign, In situ carcinoma, Invasive carcinoma	400 patches	400 [240, 80, 80]	20×/ 2048×1536
CRC [51]	Colorectal cancer and normal tissue classification	Adipose, Background, Debris, Lymphocytes, Mucus, Smooth muscle, Normal colon mucosa, Cancer-associated stroma, Colorectal adenocarcinoma epithelium	107,180 patches	107,180 [80,000, 20,000, 7180]	20×/ 224×224
BreakHis [52]	Malignant and benign tissue classification in breast cancer	Benign tumors, Malignant tumors	7909 patches	7909 [4745, 1582, 1582]	4×, 10×, 20×, 40×/ 700×460
Lymph [53]	Malignant lymph node cancer classification	Chronic lymphocytic leukemia, Follicular lymphoma, Mantle cell lymphoma	374 patches	374 [224, 75, 75]	40×/ 1388×1040
BACH [50] (WSI)	Breast cancer subtype segmentation	Normal, Benign, In situ carcinoma, Invasive carcinoma	10 WSIs	4483 [2388, 1214, 881]	10×/ 1024×1024
GlaS [54]	Gland segmentation	Benign gland, Malignant gland	165 patches	165 [68, 17, 80]	20×/ 775×522

2.1.3. General Image Dataset

We used the ImageNet dataset to evaluate the performance and efficiency of model architectures in general image classification. The ImageNet large-scale visual recognition dataset includes 1000 classes, ranging from common objects such as 'banana' to abstract concepts like 'bubble'. It contains 1,281,167 training images, 50,000 validation images, and 100,000 test images. The average image size within the dataset is approximately 469 × 387 pixels.

2.2. Response-Based Cross-Task Knowledge Distillation

The pretraining and fine-tuning strategy is a widely used approach to achieve high performance with relatively small datasets. However, existing pretraining methods are ineffective for pathological images, as described in the previous section. To overcome this limitation, we repurpose knowledge distillation to learn knowledge from unlabeled pathological images. RCKD is a novel transfer learning framework for pathological image analysis, which can be used as an alternative to the existing pretraining techniques.

Figure 1 illustrates the overall procedure of RCKD. The RCKD framework pretrains the backbone of a pathological image analysis model as a student model with the guidance of a high-performance teacher model. The teacher model is a high-performance network developed for a different task, such as nuclei segmentation, that can share knowledge with pathological image analysis tasks. First, the teacher model predicts binary nuclei segmentation maps of unlabeled pathological images. Then, the student model that combines the backbone and a segmentation head learns to output the binary nuclei segmentation map as close as possible to the output of the teacher model. At the same time as it is learning nuclei segmentation with the prediction of the teacher model as a pseudo label, the student model learns features useful for pathological image analysis. After pretraining, we remove the nuclei segmentation head from the student model, combine the pretrained backbone with a new head for the downstream task, and fine-tune it using the target data. The detailed procedure is described in Sections 2.2.1 and 2.2.2.

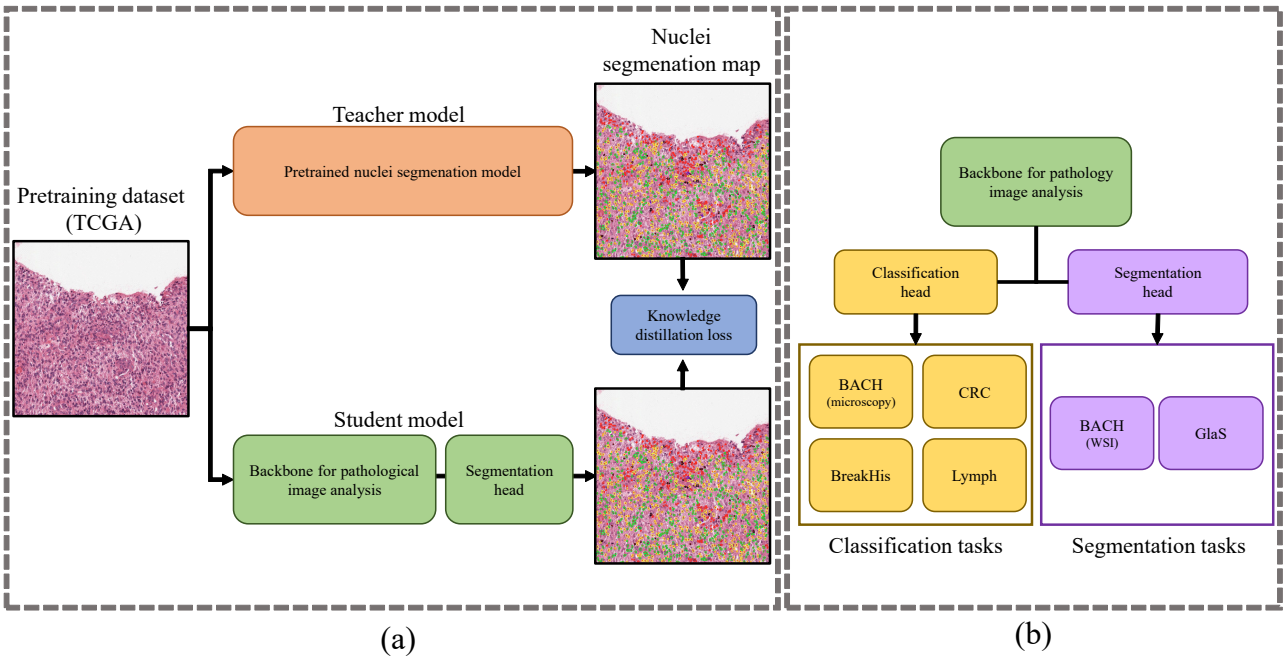

Figure 1. The overall procedure of RCKD. (**a**) Pretraining from unlabeled pathological images. (**b**) Fine-tuning for downstream tasks.

Nuclei segmentation is an effective upstream task to pretrain pathological image analysis models for the following reasons. First, pathological images are generally composed of textures rather than large-scale objects with fixed shapes. Therefore, learning low-level local features is crucial in most pathological image analysis tasks. Since nuclei segmentation is a pixel-level classification task, it drives the model to learn local and positional information. Second, a cell is the basic unit of an organ [55], and the nucleus is the core component of a cell. Most pathological images contain many nuclei and can, therefore, be used as training data for nuclei segmentation. Third, since nuclei have less shape variation than other components in pathological images, the features learned for nuclei segmentation can be useful for analyzing other types of pathological images. Fourth, the state-of-the-art (SOTA) nuclei segmentation models provide excellent performance and are sufficient to guide the pretraining of the student model.

2.2.1. Pretraining from Unlabeled Pathological Images

The teacher model takes an unlabeled pathological image x as input and predicts a binary nuclei segmentation map as Equation (1).

$$y = f_{teacher}(x), \ for \ x \in D, \tag{1}$$

where $f_{teacher}(\cdot)$ is a teacher model, $D \subset \mathbb{R}^{3xHxW}$ is a set of pathological images without segmentation labels, and $y \in \mathbb{R}^{HxW}$ is a probability map predicted by $f_{teacher}(\cdot)$, where y_{ij} is the estimated probability that a pixel x_{ij} belongs to a nuclei region. Then, we convert y into a binary segmentation map $N(x) \in \{0,1\}^{HxW}$ with a threshold value α, as Equation (2).

$$N(x)_{ij} = \begin{cases} 1 \text{ if } y_{ij} >= \alpha \\ 0 \text{ otherwise,} \end{cases} \tag{2}$$

where $\alpha = 0.5$ in our study.

For the teacher model $f_{teacher}(\cdot)$, we used StarDist [56] pretrained on the MoNuSeg2018 [57] and TNBC [58] datasets. StarDist won the CoNIC challenge in 2022 [59]. StarDist segments nuclei regions using a U-Net [60] based model and represents them as star convex polygons. The structure and hyperparameters of StarDist are presented in Figure 2. Figure 3 displays the pathological image samples used for pretraining and the pseudo label $N(x)$ estimated by StarDist.

Stage	Input size	Ouput size	Teachder model hyper-parameters
Encoder			
Stage1	512x512	256x256	L=2, D=32, $Conv_{3x3}$
Stage2	256x256	128x128	L=2, D=64, $Conv_{3x3}$
Stage3	128x128	64x64	L=2, D=128, $Conv_{3x3}$
Bridge			
Stage4	64x64	128x128	L=2, D=128, $Conv_{3x3}$
Decoder			
Stage5	128x128	256x256	L=2, D=64, $Conv_{3x3}$
Stage6	256x256	512x512	L=2, D=32, $Conv_{3x3}$
Stage7	512x512	512x512	L=2, D=32, $Conv_{3x3}$ L=1, D=128, $Conv_{3x3}$ L=1, D=1, $Conv_{1x1}$

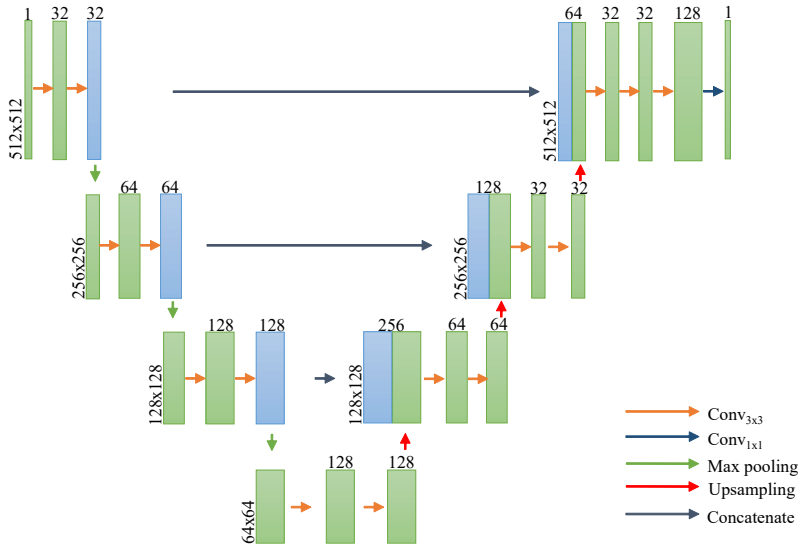

Figure 2. The structure and hyperparameters of the teacher model, StarDist. The kernel size and stride of the Max pooling are two in all layers. For upsampling, StarDist utilizes bilinear interpolation with an upsampling ratio of two.

To pretrain the backbone network via RCKD, we combine it with a segmentation head and train the combined model to predict the segmentation map identical to the pseudo label $N(x)$ for each image $x \in D$. In this study, we implemented the backbone network using a novel lightweight network, CSAT, described in Section 2.3, and the segmentation head using the U-Net decoder. The pretraining loss $L_{KD}(\theta|x)$ is defined as Equation (3).

$$\begin{aligned} L_{KD}(\theta|x) &= \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} CE(f_{student}(x_{ij};\theta), N(x_{ij})), \\ \theta &= \text{argmin}_{\theta} \frac{1}{|D|} \sum_{x \in D} L_{KD}(\theta|x) \end{aligned} \tag{3}$$

where $f_{student}(x;\theta)$ is a student model parameterized by θ and $CE(\cdot,\cdot)$ denotes the cross-entropy loss. After pretraining, we transfer the student model $f_{student}(x;\theta)$, replace the nuclei segmentation head with a new head for the target task, and fine-tune it on the target data.

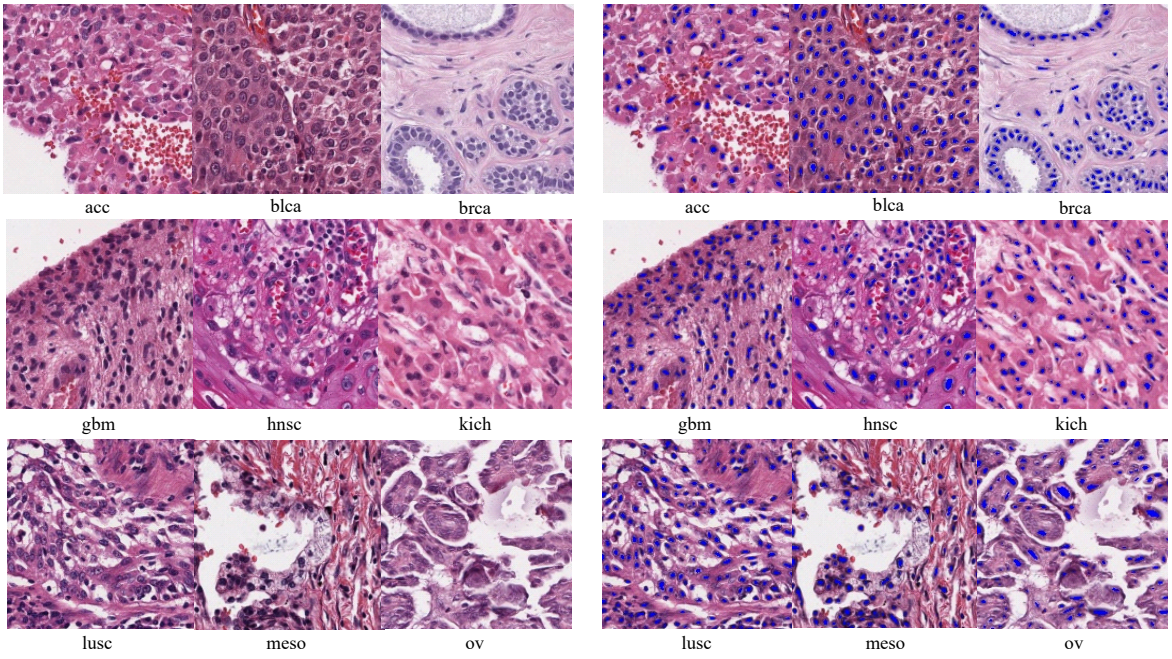

Figure 3. Pathological images in the TCGA dataset by organ (**a**) and the corresponding binary nuclei segmentation maps estimated by the teacher model, StarDist (**b**). The blue color indicates the estimated nuclei regions. (acc: adrenocortical carcinoma, blca: bladder urothelial carcinoma, brca: breast invasive carcinoma, gbm: glioblastoma multiforme, hnsc: head and neck squamous cell carcinoma, kich: kidney chromophobe, lusc: lung squamous cell carcinoma, meso: mesothelioma, and ov: ovarian serous cystadenocarcinoma).

Following the unified scheme for ImageNet (USI) [61], we first pretrain the student model on ImageNet data to improve performance and training speed before RCKD. We optimize the student model using the layer-wise adaptive rate scaling (LARS) optimizer [62] with an initial learning rate of zero and a batch size of 64. LARS automatically adjusts the learning rate for each layer, allowing training to proceed even when the initial learning rate is set to zero. We normalize the input image to 512×512 size by bilinear interpolation and warm up for 10 epochs out of 100 training epochs. The pretraining procedure is summarized in Algorithm 1.

Algorithm 1 Pretraining Procedure

W ▷ WSI in pretraining dataset (TCGA)
KD ▷ Knowledge distillation
$f_{teacher}$ ▷ Teacher model
$f_{student}$ ▷ Student model
Mag ▷ Magnification ratio of WSI
P_{size} ▷ Patch size
S_n ▷ Number of samples for pretraining
procedure PRETRAIN($W, KD, f_{teacher}, f_{student}, Mag, P_{size}, S_n$)
    $P = GetPatches(W, Mag, (P_{size}, P_{size}))$ ▷ Get patches from a WSI
    $P_{tissue} = SelectTissuePatches(P)$ ▷ Discard background patches
    $D = RandomSampling(P_{tissue}, S_n)$ ▷ Select patches for pretraining
    for all x in D **do**
        $N = f_{teacher}(x)$ ▷ Make pseudo label
        $Loss = KD(f_{student}(x), N)$ ▷ Knowledge distillation
        $Backpropagate(Loss, f_{student})$
    return $f_{student}$

* In this study, $Mag = 20$, $P_{size} = 1024$, and $S_n = 400,000$.

2.2.2. Fine-Tuning for Pathological Image Analysis Tasks

To transfer the knowledge of the pretrained student model to the downstream task, we first build a model $f_{target}(x;\theta)$ for the target task by replacing the nuclei segmentation head used to pretrain the student model with a new head for the target task. Then, we fine-tune $f_{target}(x;\theta)$ by supervised learning to minimize a task-specific loss on the labeled target dataset. In this study, we fine-tuned $f_{target}(x;\theta)$ to four classification datasets: BACH (microscopy), CRC, BreakHis, Lymph, as well as two segmentation datasets: BACH (WSI), GlaS. We used focal loss [63] to mitigate potential problems caused by data imbalance.

The model for a classification task predicts the probability that the input image belongs to each class in the form of a vector as Equation (4).

$$y_s(x) = f_{target}(x;\theta) \in \mathbb{R}^T, \tag{4}$$

where T is the number of target classes. We fine-tune the parameter θ to minimize the focal loss defined as Equation (5).

$$L_{classification}(\theta|x) = -\alpha(1 - y_s(x)_t)^\gamma \log y_s(x)_t, \tag{5}$$

where $(x,t) \in D_{target}$ is a pair of target data, and its class label and parameters α and γ are set to 0.25 and 2, respectively.

A segmentation model predicts the probability of each target class (e.g., an object or region) for each pixel in the form of a 3D map as Equation (6).

$$y_s(x) = f_{target}(x;\theta) \in \mathbb{R}^{T\times H\times W}, \tag{6}$$

where $H \times W$ is the size of the input image, and T is the number of target objects or regions. The loss function for the fine-tuning of a segmentation task is defined as Equation (7).

$$L_{segmentation}(\theta|x) = -\frac{1}{HxW}\sum_{i}^{H}\sum_{j}^{W}\alpha(1 - y_s(x)_{tij})^\gamma \log y_s(x)_{tij} \tag{7}$$

When fine-tuning the model for classification tasks, we implemented classification heads by combining a linear layer and a softmax layer. For segmentation tasks, we built segmentation heads following the decoder of U-Net. To reduce the potential bias caused by the composition of the training, validation, and test sets, which can be serious in pathological image analysis where the number of samples is usually small, we repeated the data splitting and experiment five times by changing the random seed and evaluated the models by the average performance. The fine-tuning procedure is summarized in Algorithm 2.

Algorithm 2 Fine-tuning Procedure

f_{target} ▷ Pretrained student model
D ▷ Target dataset
N_{fold} ▷ Total number of folds
N_{epoch} ▷ Total number of training epochs
N_{stop} ▷ Patience number for early stopping
procedure FINE-TUNE($f_{target}, D, N_{fold}, N_{epoch}, N_{stop}$)
  for *fold* in range [1, N_{fold}] **do**
    $X_{train}, Y_{train}, X_{val}, Y_{val}, X_{test}, Y_{test}$ = LoadData(D,*fold*)
    for *epoch* in range [1, N_{epoch}] **do**
      for i in range [1, $|X_{train}|$] **do**
        $L_{train} = L_{focal}(f_{target}(X_{train}(i)), Y_{train}(i))$
        Backpropagate(L_{train}, f_{target})
      $L_{Val} = \frac{1}{|X_{val}|)}\sum_{j=1}^{|X_{val}|} L_{focal}(f_{target}(X_{val}(j), Y_{val}(j))$
      CheckForEarlyStopping(L_{val}, N_{stop})
    $A_{test} = \frac{1}{|X_{test}|}\sum_{k=1}^{|X_{test}|}$Accuracy$(f_{target}(X_{test}(k)), Y_{test}(k))$
  $A_{average} = \frac{1}{N_{fold}}\sum_{fold=1}^{N_{fold}} A_{test}(fold)$
  return $A_{average}$

* In this study, $N_{fold} = 5$, $N_{epoch} = 200$, and $N_{stop} = 20$.

2.3. Convolutional Neural Network with Spatial Attention by Transformer

In this section, we present CSAT, a lightweight network designed for use as a general-purpose backbone network in high-resolution pathological image analysis. CSAT combines multiple techniques that have been proven effective in recent studies on lightweight networks. In addition, we improved its performance by adding a novel SAT module. Moreover, we reduce the computational and memory requirements of the Transformer for computing spatial attention by estimating attention weights at a reduced resolution and then upsampling the attention map to the size of the feature map.

2.3.1. Overall Structure

CoAtNet and AlterNet [64] exhibited improved performance by applying convolution for feature extraction in the low-level layers and aggregating features using multi-head self-attention (MSA) or Transformer blocks in the high-level layers. A few subsequent models, such as EfficientFormerV2, apply similar network configurations. CSAT also applies convolutions to extract features in the low-level layers and Transformers to aggregate feature maps in the high-level layers. The overall structure of CSAT is illustrated in Figure 4.

The structure and hyper-parameters of CSAT are based on EfficientFormerV2, a lightweight network designed for mobile environments. The bottom of CSAT consists of a patchfy stem [37] that splits the image into patches and reshapes each patch into a vector. These patches are then fed into the subsequent stages. Stages 1 and 2 are composed of two SAT blocks described in Section 2.3.2. Stages 3 and 4 consist of six and four SAT blocks followed by two Transformer blocks [65]. The SAT block extracts local features by convolutions and then re-scales the feature elements by spatial attention computed by an SAT module. The structure and hyper-parameters of CSAT are presented in Table 2.

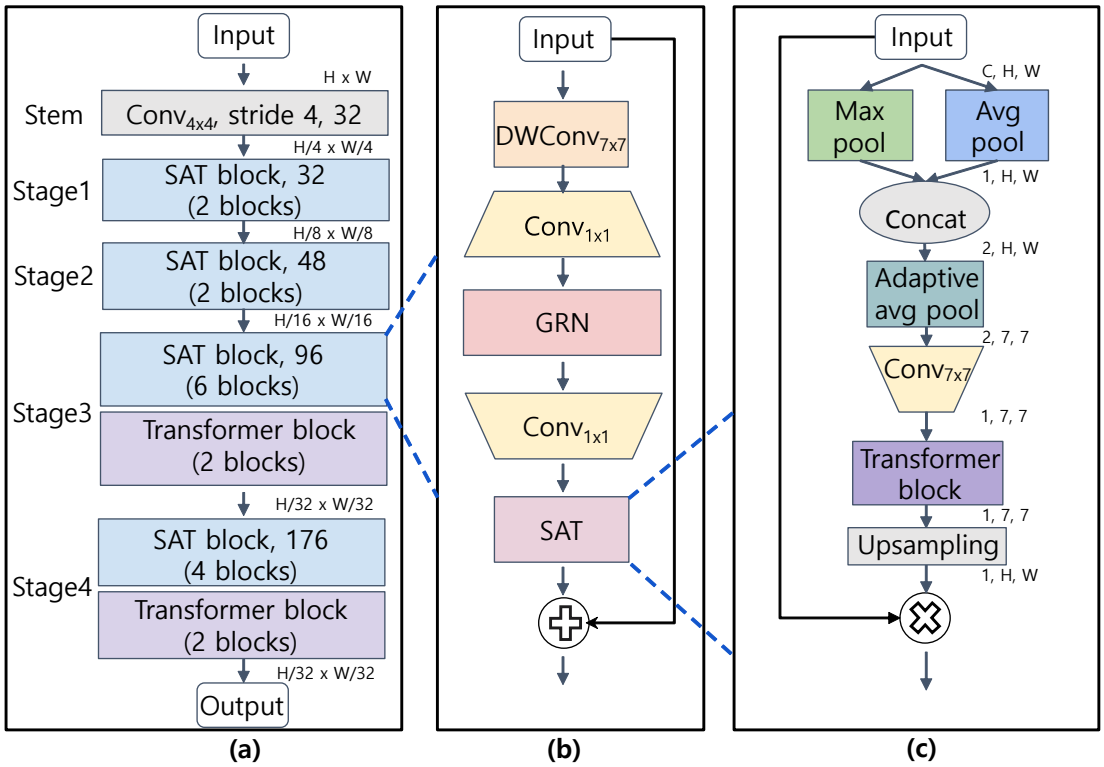

Figure 4. The architecture of CSAT. (**a**) overall structure, (**b**) SAT block, and (**c**) SAT module. $\text{Conv}_{k\times k}$, $\text{DWConv}_{k\times k}$ represent the convolution and depthwise convolution with a kernel size of $k \times k$, respectively. C, H, and W denote the channel, height, and width of the input image, respectively.

Table 2. The structure and hyper-parameters of CSAT. L denotes the number of blocks and D denotes the number of channels.

Stages	Input Size	Output Size	CSAT Hyper-Parameters
Stem	H, W	H/4, W/4	L = 1, D = 32, Convolution block
Stage1	H/4, W/4	H/8, W/8	L = 2, D = 32, SAT block
Stage2	H/8, W/8	H/16, W/16	L = 2, D = 48, SAT block
Stage3	H/16, W/16	H/32, W/32	L = 6, D = 96, SAT block L = 2, D = 96, Transformer block
Stage4	H/32, W/32	H/32, W/32	L = 4, D = 176, SAT block L = 2, D = 176, Transformer block

2.3.2. SAT Block

We designed the front part of the SAT block following Woo et al. [38]. It first extracts local features using a depthwise convolution followed by a 1×1 convolution as in Howard et al. [32]. Then, it applies global response normalization (GRN) to prevent any particular feature map from being overly dominant. It also applies an additional 1×1 convolution to abstract the feature map.

The rear part consists of a SAT module. As described in Section 2.3.3, the SAT module re-scales the features by multiplying them by attention weights estimated from the global context via a Transformer. Finally, the SAT block adds the input feature map, as in He et al. [31]. Equation (8) summarizes the operation of the SAT block.

$$y = x + SAT(Conv_{1\times1}(GRN(Conv_{1\times1}(DWConv_{7\times7}(x))))), \tag{8}$$

where $Conv_{k\times k}(\cdot)$ and $DWConv_{k\times k}(\cdot)$ denote convolution and depthwise convolution with a kernel size of $k \times k$, respectively.

2.3.3. Spatial Attention by Transformer (SAT) Module

Spatial attention is a mechanism in which the model emphasizes important features by multiplying feature elements by importance weights computed for each position. Lots of previous studies compute spatial attention weights by convolution [47,66]. However, such models have a limitation in that they estimate importance weights only from local contexts without considering the global context.

To overcome this limitation, we propose a novel SAT module that refers to the global context to compute spatial attention maps by applying a Transformer instead of convolution. In CSAT, such modification increases the number of parameters by only 1.8 K. However, the experimental results presented in Section 3.4 show that attention maps estimated from the global context can lead to higher performance than those estimated from local contexts in multiple pathological image analysis tasks.

One burden of applying a Transformer to a lightweight network is its computational and memory complexity, which scales as the square of the input resolution. To reduce the computational and memory overhead, we compute the attention map at a reduced resolution and then upsample the attention map to match the resolution of the feature maps.

Following Woo et al. [47], we first reduce the channel dimension by applying max-pooling and average-pooling in the channel direction as $x^s_{max} = P_{max}(\cdot) \in \mathbb{R}^{1\times H\times W}$, $x^s_{avg} = P_{avg}(\cdot) \in \mathbb{R}^{1\times H\times W}$ and concatenate the results as $[x^s_{max}, x^s_{avg}] \in \mathbb{R}^{2\times H\times W}$.

In general, each channel in the feature map represents a concept or object [42], and a pooling operation along the channel axis aggregates this information at each position. Then, we downsample the feature map using adaptive average pooling $P_{aap}(\cdot)$ with a fixed output resolution of $h' \times w'$, where $h' < h$ and $w' < w$. In this study, we set $h' = w' = 7$. We apply an additional convolution layer to abstract the reduced feature map.

$$M_S = f_{spatial}(x) = Conv_{7x7}(P_{aap}([P_{avg}(x); P_{max}(x)])) \tag{9}$$

The Transformer block takes as input the reduced feature map M_S combined with positional encoding and produces an attention map at a resolution of $h' \times w'$. SAT applies relative positional encoding computed by the position encoding generator (PEG) [67]. PEG encodes positional information based on the local neighborhood of input tokens, which makes the model applicable to images of different resolutions.

The Transformer block computes an attention map from the global context at a reduced scale of $h' \times w'$ as Equation (10). Since we set h' and w' to small numbers, the additional overhead is minimal.

$$\begin{aligned} Q, K, V &= Linear(M_S + PEG(M_S)) \\ SA(Q, K, V) &= softmax(\frac{QK^T}{\sqrt{d}})V, \end{aligned} \tag{10}$$

where Q, K, and V are the query, key, and value vectors, d is feature dimension, and $SA(\cdot)$ denotes self-attention.

Finally, we upsample the attention map to the size of the input feature map through bilinear interpolation and multiply the upsampled attention map to the feature map as Equation (11).

$$\begin{aligned} f_{att}(M_S) &= Upsample(SA(Q, K, V)) \\ SAT(x) &= x \times f_{att}(f_{spatial}(x)). \end{aligned} \tag{11}$$

2.4. Evaluation Metrics and Experimental Environments

We evaluated the performance in classification and segmentation tasks using the metrics of accuracy and mean intersection over union (mIoU), respectively, which are computed as Equation (12).

$$\begin{aligned} accuracy(\%) &= \frac{\text{\# of correctly classified samples}}{\text{total \# of samples}} \times 100\% \\ mIoU &= \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i + FN_i} \end{aligned} \tag{12}$$

where C, TP, FP, and FN, respectively denote the number of classes, true positives, false positives, and false negatives. Accuracy indicates how accurately the model predicts the class of the input image across the entire dataset, while mIoU represents the average of the ratio of the intersection over the union between the predicted and ground truth regions across multiple classes, measuring the overlap between them.

We conducted experiments on a computer equipped with eight NVIDIA RTX A5000 GPUs, an Intel Xeon Gold 6226R CPU, and 540 GB of RAM. We built the software environment based on Ubuntu 20.04, PyTorch v2.0 [68], CUDA v11.1, and CuDNN v8.5. We counted the parameters and measured the amount of computation of the models in MACs using the THOP library [69]. For the downstream tasks, we only counted the parameters of the backbone because the parameters and MACs vary depending on the type of task. However, in Section 3.5, we compared the number of parameters of the entire model with those of the baseline models, including the task-specific head.

3. Results

In this section, we present the results of experiments to evaluate the performance and efficiency of the proposed RCKD and CSAT compared with the pretraining methods and model architectures proposed in previous studies.

3.1. Hyperparameter Search for Downstream Tasks

We first conducted preliminary experiments on the Lymph dataset to choose hyperparameters. We applied multiple candidate values for each hyperparameter and compared the performance. For the input resolution, we compared two candidates: 224 × 224 and

384 × 384. For the learning rate, we compared 0.001 and 0.0001. We also compared the performance of the models trained with and without freezing the positional encoding.

The results are presented in Figure 5. The best performance was achieved with an input resolution of 384 × 384, a learning rate of 0.0001, and a frozen positional encoder. We used these settings in all experiments. We fine-tuned the models with stochastic gradient descent (SGD) optimizer with a batch size of 64. We trained the models for a maximum of 200 epochs. However, we stopped training if the validation loss did not decrease for more than 20 epochs.

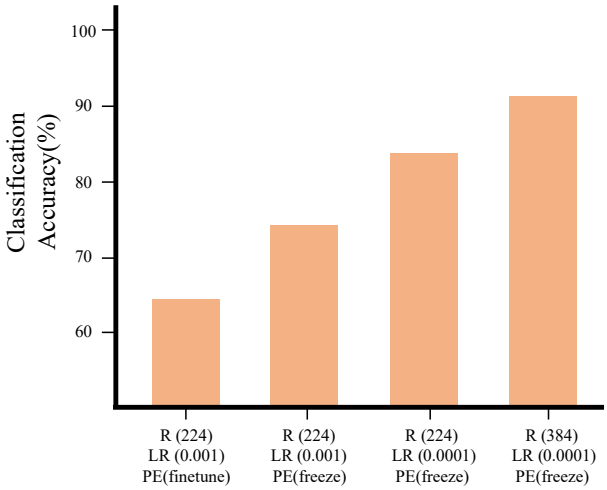

Figure 5. Classification accuracy of CSAT models on the Lymph dataset according to hyperparameters. The graph displays the average result of five experiments conducted with different data splits. R, LR, and PE denote the resolution of the input image, the learning rate, and the weight used in the positional encoding, respectively.

3.2. Performance in Downstream Tasks by Pretraining Method

We compared the performance of RCKD with three widely used pretraining methods: supervised pretraining (SPT) on the ImageNet dataset and two contrastive learning methods, Barlow Twins and MoCo. We pretrained four models for each of two distinct network architectures, ResNet18 and CSAT, employing four different pretraining methods. After transferring these pretrained models to the six downstream tasks listed in Table 1, we fine-tuned them on the target datasets. Then, we measured the performance of each model on the corresponding test sets. We also measured the performance of a model trained for the downstream tasks from random parameters without any pretraining. Figure 6 and Table 3 present the results.

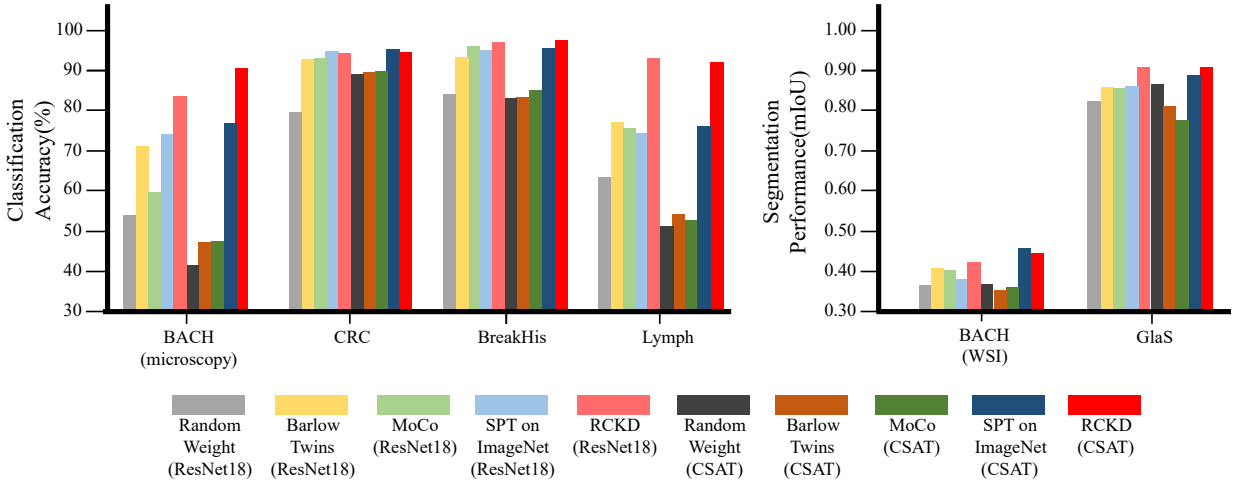

Figure 6. The performance of CSAT and ResNet18 by pretraining methods for four classification tasks (BACH (microscopy), CRC, BreakHis, and Lymph) and two segmentation tasks (BACH (WSI) and GlaS).

RCKD significantly outperformed the three baseline pretraining methods. When applied to ResNet18, RCKD exhibited an average accuracy of 92.6% in the classification tasks, which is 7.6∼11% higher than the baseline pretraining methods. When applied to CSAT, the average accuracy of RCKD was 94.2%, which is 7.9∼25.5% higher than the baseline methods. In the segmentation tasks, RCKD showed average mIoUs of 0.665 and 0.673 when applied to ResNet18 and CSAT, respectively, which are 0.035∼0.046 and 0.002∼0.107 mIoUs higher than the baseline methods.

Table 3. The performance of CSAT and ResNet18 by pretraining methods on six pathological image analysis tasks. The boldface indicates the best performance. (Params, mIoU, and GMAC denote the number of parameters, mean intersection over union, and giga multiply-accumulate operations, respectively).

Model	Params	GMAC	Pretraining Methods	Classification Accuracy (%)					Segmentation Performance (mIoU)		
				BACH (Microscopy)	CRC	BreakHis	Lymph	Average Accuracy	BACH (WSI)	GlaS	Average mIoU
ResNet18	10.6 M	5.35	Random Weight	53.7	80.2	85.0	63.7	70.6	0.355	0.83	0.592
	10.6 M	5.35	Barlow Twins	71.4	93.7	93.8	77.8	84.1	0.40	0.861	0.630
	10.6 M	5.35	MoCo	59.7	93.8	96.9	76.0	81.6	0.391	0.864	0.627
	10.6 M	5.35	SPT on ImageNet	74.4	**95.5**	95.5	74.6	85	0.373	0.866	0.619
	10.6 M	5.35	RCKD	**83.9**	95.0	**98.0**	**93.8**	**92.6**	**0.415**	**0.915**	**0.665**
CSAT	2.8 M	1.08	Random Weight	41.3	89.6	83.7	51.4	66.5	0.355	0.872	0.613
	2.8 M	1.08	Barlow Twins	47.0	90.2	83.7	53.9	68.7	0.342	0.81	0.576
	2.8 M	1.08	MoCo	47.1	90.6	85.5	52.6	68.9	0.35	0.783	0.566
	2.8 M	1.08	SPT on ImageNet	77.0	**95.8**	96.1	76.5	86.3	**0.441**	0.902	0.671
	2.8 M	1.08	RCKD	**90.6**	95.3	**98.6**	**92.5**	**94.2**	0.435	**0.912**	**0.673**

3.3. Performance in Downstream Tasks by Model Architecture

To evaluate the efficiency and effectiveness of the proposed CSAT, we conducted a comparative evaluation with two recently developed lightweight models, EfficientNet-B0 and ConvNextV2-Atto, which show outstanding performance with a small number of parameters. EfficientNet-B0 and ConvNextV2-Atto consist of 3.8 M and 3.2 M parameters, respectively, which are slightly larger than the size of CSAT with 2.8 M parameters.

Figure 7 and Table 4 present the evaluation results.

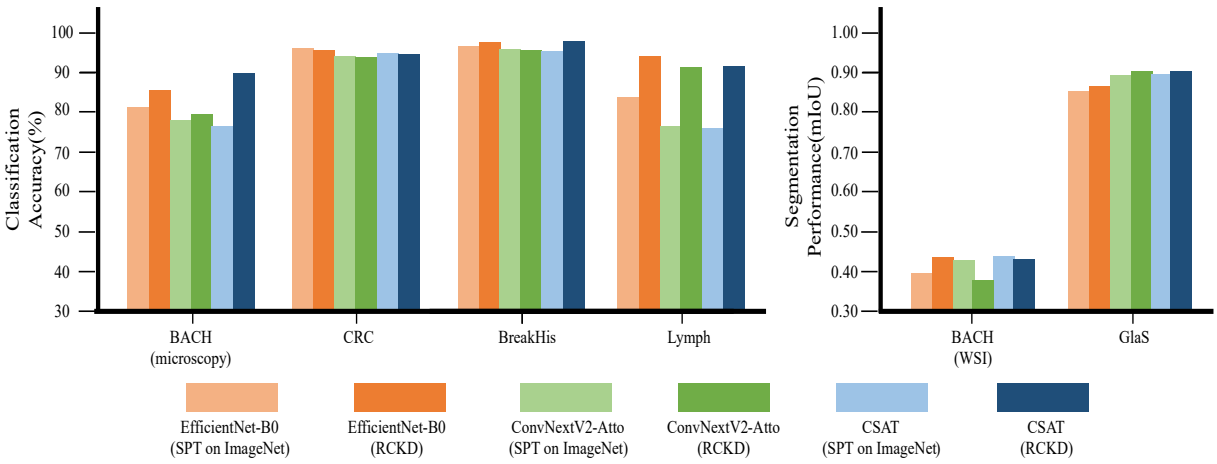

Figure 7. The performance of CSAT and recent lightweight models on six pathological image analysis tasks. (SPT denotes supervised pretraining).

In the classification tasks, the CSAT pretrained by RCKD showed the best average accuracy of 94.2%. This model performed best for the BACH and BreakHis datasets. The supervised pre-

trained EfficientNet-B0 performed best on the CRC dataset, and the EfficientNet-B0 pretrained by RCKD performed best on the Lymph dataset. Despite having only 2.8 M parameters, which is only 73.6% of the parameters in EfficientNet-B0, CSAT showed slightly higher performance on average. Among the supervised pretrained baseline models, EfficientNet-B0 exhibited the highest average accuracy of 90.2%. CSAT pretrained by RCKD outperformed this model by 4%.

Table 4. The performance of CSAT and recent lightweight models on six pathological image analysis tasks. The boldface indicates the best performance. (Params, mIoU, and GMAC denote the number of parameters, mean intersection over union, giga multiply-accumulate operations, respectively).

Model	Params	GMAC	Pretraining Methods	Classification Accuracy (%)					Segmentation Performance (mIoU)		
				BACH (Microscopy)	CRC	BreakHis	Lymph	Average Accuracy	BACH (WSI)	GlaS	Average mIoU
EfficientNet-B0	3.8 M	1.21	SPT on ImageNet	82.2	**96.8**	97.4	84.7	90.2	0.399	0.861	0.630
EfficientNet-B0	3.8 M	1.21	RCKD	85.9	96.2	98.3	**94.9**	93.8	0.438	0.873	0.655
ConvNextV2-Atto	3.2 M	1.60	SPT on ImageNet	78.6	95.1	96.6	77.2	86.8	0.434	0.901	0.667
ConvNextV2-Atto	3.2 M	1.60	RCKD	80.4	94.9	96.4	91.9	90.9	0.38	0.91	0.645
CSAT	2.8 M	1.08	SPT on ImageNet	77.0	95.8	96.1	76.5	86.3	**0.441**	0.902	0.671
CSAT	2.8 M	1.08	RCKD	**90.6**	95.3	**98.6**	92.5	**94.2**	0.435	**0.912**	**0.673**

CSAT pretrained by RCKD also performed best in the segmentation tasks, showing an average mIoU of 0.673. However, on the BACH (WSI) dataset, the supervised pretrained CSAT exhibited the best performance. Both CSAT models outperformed EfficientNet-B0 and ConvNextV2-Atto on average.

3.4. Comparison with Previous Studies on Pathological Image Analysis

We compared the proposed methods with two recent studies on pathological image analysis, Riasatian et al. [70] and Ciga et al. [12]. Riasatian et al. [70] present KimiaNet pretrained by weakly supervised learning. Ciga et al. [12] apply SimCLR, a contrast learning method for pretraining. We also included the masked auto-encoder (MAE) [71] in the baseline models because, although it was not specialized for pathological image analysis, it has shown excellent performance in recent studies on computer vision. For Riasatian et al. [70] and Ciga et al. [12], we initialized the model by the pretrained parameters provided by the authors. However, because we were unable to find any pretrained MAE models specifically designed for pathological images, we only pretrained the MAE model using the TCGA dataset.

Figure 8 and Table 5 present the model architecture, pretraining method, the number of parameters, computational complexity in GMAC, and the performance of the models.

Table 5. The performance of CSAT pretrained by RCKD compared with three previous studies on six pathological image analysis tasks. The boldface indicates the best performance. (Params, mIoU, and GMAC denote the number of parameters, mean intersection over union, and giga multiply-accumulate operations, respectively).

Pretraining Methods	Model	Params	GMAC	Classification Accuracy (%)					Segmentation Performance (mIoU)		
				BACH (Microscopy)	CRC	BreakHis	Lymph	Average Accuracy	BACH (WSI)	GlaS	Average mIoU
KimiaNet [70]	DensNet121	6.6M	8.51	61.3	93.8	96.3	71.5	80.7	0.385	0.804	0.594
SimCLR [12]	ResNet18	10.6M	5.35	76.2	94.2	97.7	**93.3**	90.3	0.386	0.866	0.626
MAE [71]	ViT-B	81.6M	49.3	62.3	94.1	93.1	71.3	80.2	0.378	0.763	0.570
RCKD	CSAT	2.8M	1.08	**90.6**	**95.3**	**98.6**	92.5	**94.2**	**0.435**	**0.912**	**0.673**

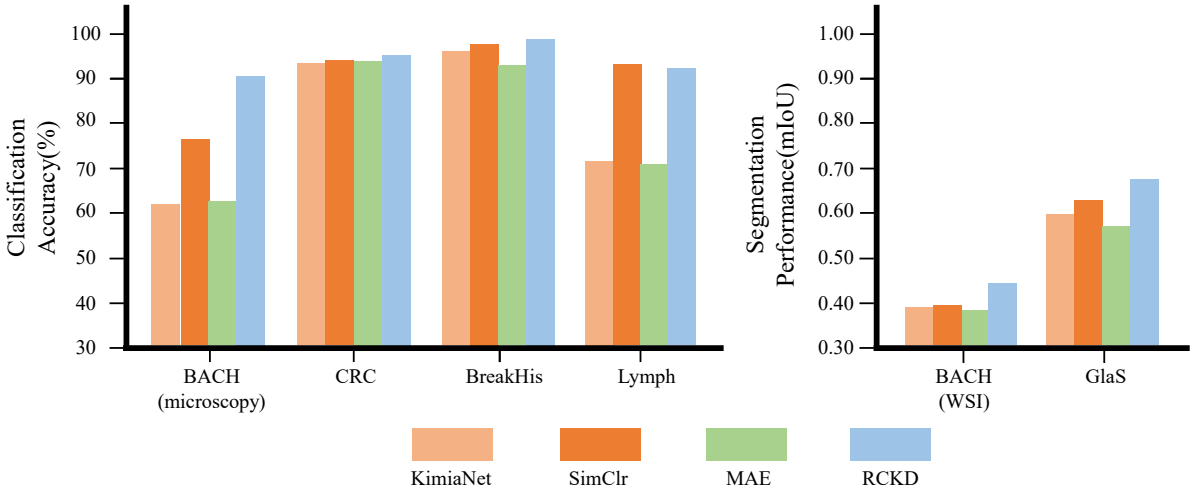

Figure 8. The performance of CSAT pretrained by RCKD compared with three previous studies on six pathological image analysis tasks.

Despite the significantly smaller number of parameters and lower computational complexity compared to the baseline models, the proposed model performed best on five datasets among six downstream datasets and outperformed the baseline models on average in both classification and segmentation tasks. On the Lymph dataset, ResNet18 pretrained with SimCLR performed best. However, the proposed model showed a classification accuracy on the Lymph dataset only 0.8% lower than the best model [12] with only 26.4% of the parameters and 20.1% of the computation compared to the best model.

3.5. Evaluation of CSAT on ImageNet

Finally, we evaluated the performance of CSAT in general image classification using the ImageNet dataset. In this experiment, we compared CSAT with ResNet18 because ResNet18 is a lightweight model and widely used in computer vision. For an ablation study, we built two variants for both CSAT and ResNet18, one with the SAT module applied and one without. We trained the four models following USI [61], which is based on knowledge distillation and modern tricks. Figure 9 and Table 6 presents the results. With the SAT module, CSAT showed 2.4% higher classification accuracy than ResNet18 combined with the SAT module using 73.8% fewer parameters and 79.9% less computation. Without the SAT module, CSAT exhibited 2.7% higher accuracy than the vanilla ResNet18 model. The SAT module increased the accuracy of ResNet18 by 0.5% and that of CSAT by 0.2%.

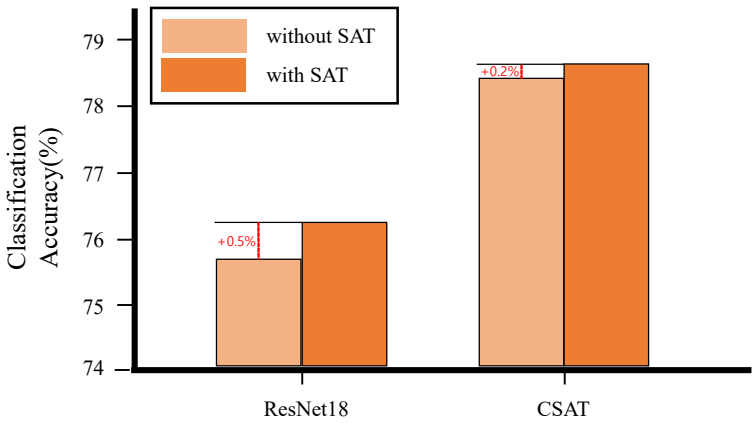

Figure 9. The performance of CSAT and ResNet18 on the ImageNet dataset.

Table 6. The performance of CSAT and ResNet18 on the ImageNet dataset. (Params and GMAC represent the number of parameters and giga multiply-accumulate operations, respectively).

Model	SAT Module	Params	GMAC	Classification Accuracy (%)
ResNet18	X	11,689,512	5.35	75.7
ResNet18	O	11,690,544	5.35	76.2
CSAT	X	3,063,272	1.08	78.4
CSAT	O	3,065,078	1.08	78.6

3.6. Accuracy vs. Efficiency

In order to effectively analyze high-resolution pathological images in a general environment, not only performance but also computational and parameter efficiency are important. Therefore, we compared the accuracy, the amount of computation, and the number of parameters of models according to architecture and pretraining method. Figure 10 summarizes the results. The horizontal axis represents the amount of computation, while the vertical axis represents the average classification accuracy over the BACH (microscopy), CRC, BreakHis, and Lymph datasets. The color and size of each circle represent the pretraining methods and the number of parameters, respectively.

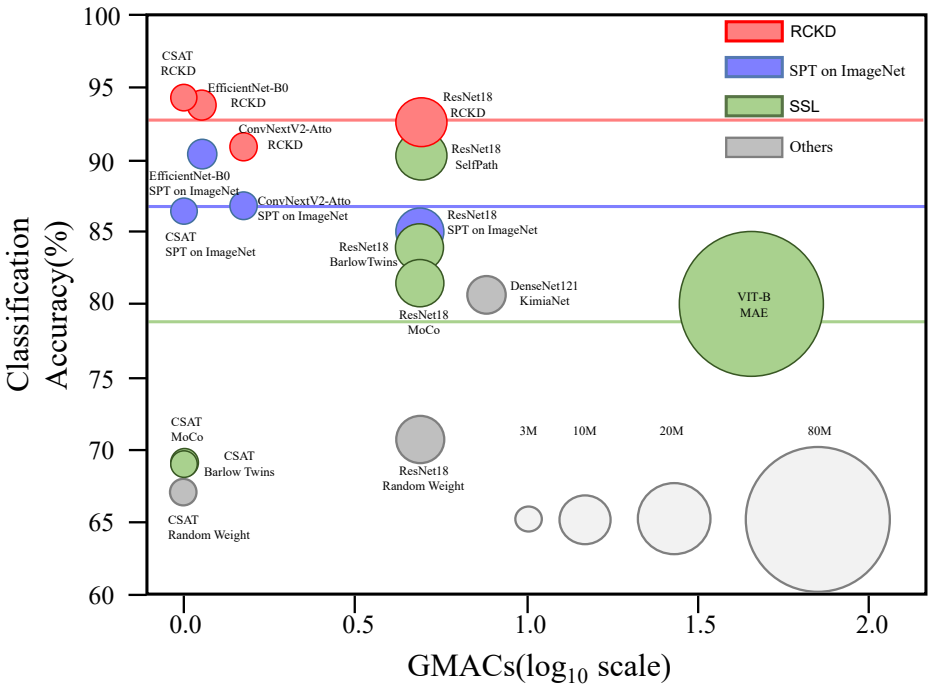

Figure 10. A ball chart displaying the average classification accuracy according to model architecture and pretraining method. The vertical axis represents the accuracy averaged for four classification tasks listed in Table 1. The horizontal axis represents the computational complexity in GMAC. The color and size of each circle represent the pretraining methods and the number of parameters, respectively. The average classification accuracy of the models pretrained by RCKD, supervised pretraining (SPT) on ImageNet, and self-supervised learning (SSL) on pathological images are 92.8% (red line), 87% (blue line), and 78.9% (green line), respectively.

4. Discussion

As a transfer learning framework, RCKD has the following advantages. RCKD does not require labeled data for pretraining and thus enables the student model to learn a lot of knowledge from a large amount of unlabeled data. Furthermore, in RCKD, the student model learns not only from the data but also from the teacher model. This is an important advantage over conventional pretraining approaches where the model only learns from one knowledge source, the training data. Unlike conventional knowledge distillation, RCKD

can transfer knowledge from a teacher model developed for a different task. Moreover, unlike conventional transfer learning methods, RCKD learns knowledge from the teacher model without transferring model parameters, so it is applicable even when the teacher and student models have different architectures.

On the other hand, RCKD shares the limitation of KD that the performance of the student model depends on that of the teacher model. The experimental results in Section 3 show that the nuclei segmentation model, StarDist, is strong enough to guide the training of the student model. However, to apply RCKD to other fields, it is necessary to search for teacher models that perform well in the areas where knowledge can be shared with the downstream task.

It is unclear why contrastive learning and MAE are less effective in pathological image analysis than in other computer vision fields. One possible reason is the unique characteristics of pathological images that differ from general images. While general images mainly comprise objects with consistent large-scale shapes, pathological images are composed of tissues with irregular sizes and shapes. We suspect that conventional pretraining techniques prioritize global patterns over local details, despite the latter's significance in the analysis of pathological images.

5. Conclusions

Major challenges in pathological image analysis include the scarcity of labeled data and the characteristics of pathological images significantly different from ordinary images, which limits the effect of conventional transfer learning techniques. To overcome these limitations, we proposed a novel Response-based Cross-task Knowledge Distillation (RCKD) framework that learns knowledge from unlabeled pathological images guided by a teacher model developed for a different task. In experiments, RCKD outperformed supervised pretraining and contrastive learning by large margins. RCKD has additional advantages in that it does not require manual labeling and can learn knowledge from a teacher model with different architecture or target tasks. We also proposed the Convolutional neural network with Spatial Attention by Transformers (CSAT), a lightweight architecture for the processing of high-resolution pathological images, such as pathological images. CSAT outperformed ResNet18 on ImageNet by a large margin. The CSAT pretrained by RCKD exhibited average performances of 94.2% in classification tasks and 0.673 mIoU in segmentation tasks, which are 3.9∼14% and 0.047∼0.103 mIoU higher than recent pathological image analysis models, respectively. We expect that the results of this study will improve the performance and efficiency of deep learning-based pathological image analysis models, thereby accelerating the development of key techniques for AI-assisted, or fully automated diagnosis.

Author Contributions: Conceptualization, H.K. and T.-Y.K.; methodology, H.K.; software, H.K.; validation, H.K. and H.C.; formal analysis, H.K.; investigation, H.K.; resources, H.K.; data curation, H.K.; writing—original draft preparation, H.K.; writing—review and editing, I.K.; visualization, H.K.; supervision, T.-Y.K. and I.K.; project administration, T.-Y.K.; funding acquisition, S.W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1137).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the reason that this study was retrospectively performed on the publicly opened extant data (TCGA dataset) which was fully anonymized before the study was planned and designed.

Informed Consent Statement: Informed consent is known to be obtained from all subjects at the time of the study material (TCGA dataset) construction.

Data Availability Statement: The code in this article cannot be published due to privacy and can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest: Hyunil Kim and Hyeyoon Chang are employees of Deep Bio Inc. Tae-Yeong Kwak is the CTO of Deep Bio Inc., and Sun Woo Kim is the CEO of Deep Bio Inc. The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CSAT	Convolutional neural network with Spatial Attention by Transformer
CFC	Channel-wise Fully Connected
GRN	Global Response Normalization
KD	Knowledge Distillation
MAC	Multiply ACcumulate
MAE	Masked Auto Encoder
RCKD	Response based Cross task Knowledge Distillation
PEG	Position Encoding Generator
RKD	Response-based Knowledge Distillation
SAT	Spatial Attntion by Transformer
SGD	Stochastic Gradient Descent
SOTA	State Of The Art
SPT	Supervised PreTraining
SSL	Self Supervised Learning
USI	Unified Scheme for ImageNet
ViT	Vision Transformer
WSI	Whole Slide Imaging

Appendix A

This appendix provides the details of the TCGA dataset used for pretraining in Section 2.1.1.

Appendix A.1. Detail of the TCGA Dataset

In this paper, we use the TCGA dataset to validate the performance of our proposed training method. Table A1 shows the number of WSIs and image patches of the 32 types of cancer data we used.

Table A1. Total image patch of TCGA dataset extracted from WSI.

Study Abbreviation	Study Name	WSI	# of Patches	Magnification
ACC	Adrenocortical carcinoma	227	246781	20×
BLCA	Bladder Urothelial Carcinoma	458	457819	20×
BRCA	Breast invasive carcinoma	1129	769008	20×
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	278	180447	20×
CHOL	Cholangiocarcinoma	39	47288	20×
COAD	Colon adenocarcinoma	441	243863	20×
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	44	27435	20×
ESCA	Esophageal carcinoma	158	119497	20×
GBM	Glioblastoma multiforme	860	518580	20×
HNSC	Head and Neck squamous cell carcinoma	468	326319	20×
KICH	Kidney Chromophobe	121	107381	20×
KIRC	Kidney renal clear cell carcinoma	519	454207	20×
KIRP	Kidney renal papillary cell carcinoma	297	226632	20×

Table A1. *Cont.*

Study Abbreviation	Study Name	WSI	# of Patches	Magnification
LGG	Brain Lower Grade Glioma	843	549297	20×
LIHC	Liver hepatocellular carcinoma	372	320796	20×
LUAD	Lung adenocarcinoma	531	397341	20×
LUSC	Lung squamous cell carcinoma	512	394099	20×
MESO	Mesothelioma	79	52186	20×
OV	Ovarian serous cystadenocarcinoma	107	98306	20×
PAAD	Pancreatic adenocarcinoma	209	170715	20×
PCPG	Pheochromocytoma and Paraganglioma	194	182398	20×
PRAD	Prostate adenocarcinoma	450	365360	20×
READ	Rectum adenocarcinoma	157	67092	20×
SARC	Sarcoma	726	661662	20×
SKCM	Skin Cutaneous Melanoma	476	396349	20×
STAD	Stomach adenocarcinoma	400	297018	20×
TGCT	Testicular Germ Cell Tumors	211	207681	20×
THCA	Thyroid carcinoma	518	445611	20×
THYM	Thymoma	180	173342	20×
UCEC	Uterine Corpus Endometrial Carcinoma	545	585862	20×
UCS	Uterine Carcinosarcoma	87	94565	20×
UVM	Uveal Melanoma	80	44387	20×

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105 . [CrossRef]
2. Cireşan, D.; Giusti, A.; Gambardella, L.; Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, 22–26 September 2013, Proceedings, Part II 16*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 411–418.
3. Veta, M.; Van Diest, P.J.; Willems, S.M.; Wang, H.; Madabhushi, A.; Cruz-Roa, A.; Gonzalez, F.; Larsen, A.B.L.; Vestergaard, J.S.; Dahl, A.B.; et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.* **2015**, *20*, 237–248. [CrossRef] [PubMed]
4. Araújo, T.; Aresta, G.; Castro, E.; Rouco, J.; Aguiar, P.; Eloy, C.; Polónia, A.; Campilho, A. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE* **2017**, *12*, e0177544. [CrossRef] [PubMed]
5. Chen, H.; Qi, X.; Yu, L.; Heng, P. DCAN: Deep contour-aware networks for accurate gland segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2487–2496.
6. Serag, A.; Ion-Margineanu, A.; Qureshi, H.; McMillan, R.; Saint Martin, M.; Diamond, J.; O'Reilly, P.; Hamilton, P. Translational AI and deep learning in diagnostic pathology. *Front. Med.* **2019**, *6*, 185. [CrossRef]
7. Deng, S.; Zhang, X.; Yan, W.; Chang, E.; Fan, Y.L.M.; Xu, Y. Deep learning in digital pathology image analysis: A survey. *Front. Med.* **2020**, 14, 470–487. [CrossRef]
8. Khan, S.; Islam, N.; Jan, Z.; Din, I.; Rodrigues, J. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognit. Lett.* **2019**, *125*, 1–6. [CrossRef]
9. Mormont, R.; Geurts, P.; Marée, R. Comparison of deep transfer learning strategies for digital pathology. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2262–2271.
10. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
11. Boyd, J.; Liashuha, M.; Deutsch, E.; Paragios, N.; Christodoulidis, S.; Vakalopoulou, M. Self-supervised representation learning using visual field expansion on digital pathology. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 639–647.

12. Ciga, O.; Xu, T.; Martel, A. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **2022**, *7*, 100198. [CrossRef]
13. Dehaene, O.; Camara, A.; Moindrot, O.; de Lavergne, A.; Courtiol, P. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv* **2020**, arXiv:2012.03583.
14. Zhang, L.; Amgad, M.; Cooper, L. A Histopathology Study Comparing Contrastive Semi-Supervised and Fully Supervised Learning. *arXiv* **2021**, arXiv:2111.05882.
15. Li, J.; Lin, T.; Xu, Y. Sslp: Spatial guided self-supervised learning on pathological images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part II 24*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 3–12.
16. Koohbanani, N.; Unnikrishnan, B.; Khurram, S.; Krishnaswamy, P.; Rajpoot, N. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Trans. Med. Imaging* **2021**, *40*, 2845–2856. [CrossRef]
17. Lin, T.; Yu, Z.; Xu, Z.; Hu, H.; Xu, Y.; Chen, C. SGCL: Spatial guided contrastive learning on whole-slide pathological images. *Med. Image Anal.* **2023**, *89*, 102845. [CrossRef] [PubMed]
18. Tomasev, N.; Bica, I.; McWilliams, B.; Buesing, L.; Pascanu, R.; Blundell, C.; Mitrovic, J. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? In Proceedings of the First Workshop on Pre-Training: Perspectives, Pitfalls, and Paths Forward at ICML, Baltimore, MD, USA, 23 July 2022.
19. Weinstein, J.; Collisson, E.; Mills, G.; Shaw, K.; Ozenberger, B.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
20. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual Event, 12–18 July 2020; pp. 1597–1607.
21. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
22. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
23. Adriana, R.; Nicolas, B.; Ebrahimi, K.; Antoine, C.; Carlo, G.; Yoshua, B. Fitnets: Hints for thin deep nets. *Proc. ICLR* **2015**, *2*, 3.
24. Gou, J.; Yu, B.; Maybank, S.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
25. Li, D.; Wu, A.; Han, Y.; Tian, Q. Prototype-guided Cross-task Knowledge Distillation for Large-scale Models. *arXiv* **2022**, arXiv:2212.13180.
26. DiPalma, J.; Suriawinata, A.; Tafe, L.; Torresani, L.; Hassanpour, S. Resolution-based distillation for efficient histology image classification. *Artif. Intell. Med.* **2021**, *119*, 102136. [CrossRef]
27. Javed, S.; Mahmood, A.; Qaiser, T.; Werghi, N. Knowledge Distillation in Histology Landscape by Multi-Layer Features Supervision. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2037–2046. [CrossRef]
28. Zhang, R.; Zhu, J.; Yang, S.; Hosseini, M.; Genovese, A.; Chen, L.; Rowsell, C.; Damaskinos, S.; Varma, S.; Plataniotis, K. HistoKT: Cross Knowledge Transfer in Computational Pathology. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1276–1280.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
30. Iandola, F.; Han, S.; Moskewicz, M.; Ashraf, K.; Dally, W.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 207–212.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
34. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
35. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 10096–10106.
36. Dai, Z.; Liu, H.; Le, Q.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
37. Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
38. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.; Xie, S. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142.
39. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949.

40. Li, Y.; Hu, J.; Wen, Y.; Evangelidis, G.; Salahi, K.; Wang, Y.; Tulyakov, S.; Ren, J. Rethinking Vision Transformers for MobileNet Size and Speed. In Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 16889–16900.
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
44. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3024–3033.
45. Lee, H.; Kim, H.; Nam, H. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1854–1862.
46. Park, J.; Woo, S.; Lee, J.; Kweon, I. Bam: Bottleneck attention module. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; p. 147.
47. Woo, S.; Park, J.; Lee, J.; Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
48. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
49. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021. [CrossRef]
50. Aresta, G.; Araújo, T.; Kwok, S.; Chennamsetty, S.; Safwan, M.; Alex, V.; Marami, B.; Prastawa, M.; Chan, M.; Donovan, M.; et al. Bach: Grand challenge on breast cancer histology images. *Med. Image Anal.* **2019**, *56*, 122–139. [CrossRef]
51. Kather, J.; Halama, N.; Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo* **2018**, *10*, 5281.
52. Spanhol, F.; Oliveira, L.; Petitjean, C.; Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 1455–1462. [CrossRef]
53. Orlov, N.; Chen, W.; Eckley, D.; Macura, T.; Shamir, L.; Jaffe, E.; Goldberg, I. Automatic classification of lymphoma images with transform-based global features. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 1003–1013. [CrossRef]
54. Sirinukunwattana, K.; Pluim, J.; Chen, H.; Qi, X.; Heng, P.; Guo, Y.; Wang, L.; Matuszewski, B.; Bruni, E.; Sanchez, U.; et al. Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Anal.* **2017**, *35*, 489–502. [CrossRef]
55. Mason, K.; Losos, J.; Singer, S.; Raven, P.; Johnson, G. *Biology*; McGraw-Hill Education: New York, NY, USA, 2017.
56. Schmidt, U.; Weigert, M.; Broaddus, C.; Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018, Proceedings, Part II 11*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 265–273.
57. Kumar, N.; Verma, R.; Anand, D.; Zhou, Y.; Onder, O.; Tsougenis, E.; Chen, H.; Heng, P.; Li, J.; Hu, Z.; et al. A multi-organ nucleus segmentation challenge. *IEEE Trans. Med. Imaging* **2019**, *39*, 1380–1391. [CrossRef]
58. Naylor, P.; Laé, M.; Reyal, F.; Walter, T. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Trans. Med. Imaging* **2018**, *38*, 448–459. [CrossRef]
59. Graham, S.; Jahanifar, M.; Vu, Q.; Hadjigeorghiou, G.; Leech, T.; Snead, D.; Raza, S.; Minhas, F.; Rajpoot, N. Conic: Colon nuclei identification and counting challenge 2022. *arXiv* **2021**, arXiv:2111.14485.
60. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
61. Ridnik, T.; Lawen, H.; Ben-Baruch, E.; Noy, A. Solving imagenet: A unified scheme for training any backbone to top results. *arXiv* **2022**, arXiv:2204.03475.
62. You, Y.; Gitman, I.; Ginsburg, B. Large batch training of convolutional networks. *arXiv* **2017**, arXiv:1708.03888.
63. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
64. Park, N.; Kim, S. How do vision transformers work? In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022. [CrossRef]
65. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 30. [CrossRef]
66. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.

67. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. In Proceedings of the ICLR 2023, Kigali, Rwanda, 1–5 May 2023. [CrossRef]
68. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. NIPS-W. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
69. THOP: PyTorch-OpCounter. Available online: https://github.com/Lyken17/pytorch-OpCounter (accessed on 3 August 2023).
70. Riasatian, A.; Babaie, M.; Maleki, D.; Kalra, S.; Valipour, M.; Hemati, S.; Zaveri, M.; Safarpoor, A.; Shafiei, S.; Afshari, M.; et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Med. Image Anal.* **2021**, *70*, 102032. [CrossRef] [PubMed]
71. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16000–16009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.