IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 September 2023, accepted 2 October 2023, date of publication 19 October 2023, date of current version 24 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324383

RESEARCH ARTICLE

CB-HVT Net: A Channel-Boosted Hybrid Vision Transformer Network for Lymphocyte Detection in Histopathological Images

MOMINA LIAQAT ALI[1], **ZUNAIRA RAUF**[1,2], **ASIFULLAH KHAN**[1,2,3], **ANABIA SOHAIL**[1,4], **RAFI ULLAH**[5], **AND JEONGHWAN GWAK**[6]

[1]Pattern Recognition Laboratory, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad 45650, Pakistan
[2]PIEAS Artificial Intelligence Center (PAIC), Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad 45650, Pakistan
[3]Center for Mathematical Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad 45650, Pakistan
[4]Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[5]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS (UTP), Seri Iskandar, Perak 31750, Malaysia
[6]Department of Software, Korea National University of Transportation, Chungju 27469, Republic of Korea

Corresponding authors: Asifullah Khan (asif@pieas.edu.pk) and Jeonghwan Gwak (jgwak@ut.ac.kr)

This research was supported by ''Regional Innovation Strategy (RIS)'' through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-001).

ABSTRACT Detection of Tumor-Infiltrating Lymphocytes (TILs) has a high prognostic value in cancer diagnosis due to their ability to identify and kill cancer cells. However, this task is non-trivial due to their diverse morphology, overlapping boundaries, and presence of artifacts. Vision Transformers (ViTs) have the ability to capture long-range relationships, but they lack local correlation in the images and require large training datasets. In this work, we propose a Channel Boosted Hybrid Vision Transformer (CB-HVT) to detect lymphocytes in histopathological images. The proposed network constitutes: 1) channel generation module; 2) channel exploitation module; 3) channel merging module; 4) region-aware module; and 5) detection and segmentation head. The proposed CB-HVT exploits the learning capacity of both CNN and ViT-based architectures to capture lymphocytic diverse morphology. In addition, we developed a feature fusion block to systematically and gradually merge the diverse feature maps to improve the learning capability of the network. The attention mechanism in the fusion block retains the most contributing features. We evaluated the effectiveness of the proposed CB-HVT on two publicly available datasets for lymphocyte detection in histopathological images. The proposed network showed good results as compared to the existing architectures in terms of F-Score (LYSTO: 0.88 and NuClick: 0.82). In addition, the performance of the proposed CB-HVT on an unseen test set reveals its significance as a valuable tool for pathologists for real-time lymphocyte detection.

INDEX TERMS Attention, channel boosting, channel generation, CNNs, feature fusion, lymphocyte detection, transfer learning, vision transformers.

I. INTRODUCTION

Cancer is one of the deadliest diseases worldwide, causing millions of people to die each year [1], [2]. The tumor microenvironment (TME) contains specialized immune cells, known as, tumor-infiltrating lymphocytes (TILs) that play an important role in killing the cancer cells [3]. Therefore, these cells have a high clinical importance and are considered an important prognostic measure for cancer analysis [4]. Several studies have revealed the importance of TILs evaluation for the analysis of cancer progression and the therapeutic efficacy of treatments such as chemotherapy or

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Loconsole.

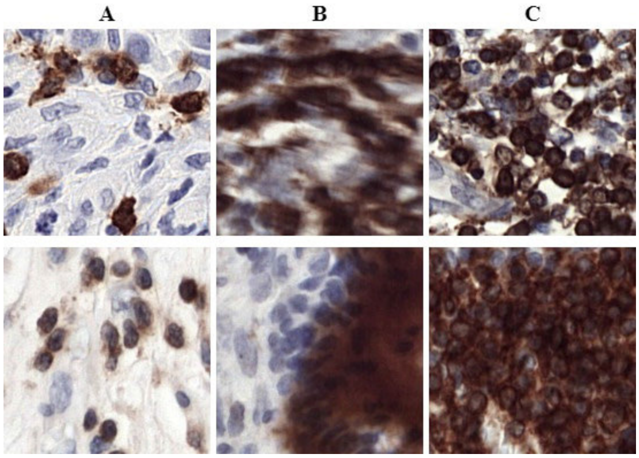

FIGURE 1. **High-level pattern diversity in IHC-stained histopathological images from LYSTO (1st row), and NuClick (2nd row) datasets. Panel A shows normally distributed lymphocytes, panel B shows the artifact regions and panel C shows regions with overlapping lymphocytes. Additionally, images from panel B (Artifacts) and panel C (crowded lymphocytes) show a high resemblance which makes lymphocyte detection challenging.**

surgery [5]. Consequently, their detection and quantification are important to measure the extent of immune response in patients during the cancer diagnosis [6].

However, in a clinical workflow, tissue slides are observed manually which is very tedious, prone to errors, and may suffer from inter- and intra-observer variability among pathologists [7]. Therefore, an accurate automated diagnostic system for lymphocyte detection can help pathologists perform accurate cancer diagnosis and devise treatment plans [8].

However, automated lymphocyte detection poses a number of challenges, including their complex morphology, presence of noise, overlapping cells, unclear boundaries, and a limited amount of pathologist-labeled datasets [9]. In addition, these tissue samples exhibit a high level of pattern diversity especially between the regions with overlapping instances and artifacts (Figure 1).

Recent advancements in deep learning have revolutionized computer vision and facilitated the development of computer-aided diagnostic systems [10]. Convolutional Neural Network (CNN) based automated systems have been developed for various tasks, including tumor classification [11], nuclei segmentation [12], COVID detection [13], and cancer analysis [14], because of their ability to learn discriminant features automatically from images [15]. However, CNNs have the limitation of focusing only on local aspects of images, which means they fail to capture the global perspective of images [16]. The small receptive fields of convolution filters (usually 3 × 3 or 5 × 5) usually capture local correlations in the images but they may fail to capture global-level information. Although many approaches have utilized dilated convolutions, large filters, and attention mechanisms to increase the receptive field, they still fall short of capturing the global perspective of images [17], [18]. Therefore, CNN-based systems that solely emphasize local patterns may demonstrate inadequate performance due to the presence of intricate multi-level complex patterns dispersed globally within the medical images [19].

Recently, vision transformers (ViTs) have gained popularity due to their ability to model long-range dependencies in the images by utilizing their multi-head self-attention mechanism and positional embeddings [20], [21], [22]. However, the poor image-related inductive bias of ViTs, high memory consumption due to multi-head attention, and fixed-sized image tokens may limit their performance for medical images [23], [24], [25]. In addition, ViT-based networks assume identical distributions for both training and test sets, however, in medical images, varying staining techniques, data acquisition methods, and scanners can introduce a domain shift, which may lead to poor generalization [20], [26], [27].

Therefore, in this study, we present a hybrid approach that leverages the benefits of both CNNs and ViTs for histopathological images. The proposed Channel Boosted Hybrid Vision Transformer network "CB-HVT Net" utilizes three main modules to learn highly correlated and domain-specific features, including a) the channel generator module, b) the channel exploitation module, and c) the channel merging module. In the channel generation module, we use three different channel generators to learn diverse channels from images, with each generator specialized in learning different types of features. By combining the knowledge space of all these generators, we achieve a boosted channel space, which undergoes effective channel exploitation and systematic channel fusion in the channel exploitation, and the channel merging modules, respectively. Later the proposed CB-HVT Net utilizes its region aware module and detection and segmentation head to generate the final output. In this regard, the proposed approach "CB-HVT" enables a more reliable and efficient automated evaluation of lymphocytes, ultimately leading to better diagnosis and treatments for patients. Figure 2 shows the detailed workflow of the proposed framework.

The significant contributions of the proposed CB-HVT Net are listed below:

- The proposed Channel Boosted Hybrid Vision Transformer Network "CB-HVT Net" integrates CNNs and ViT-based channel generators using the idea of channel-boosting. The generated diverse channels approach in the channel generation module effectively captures both local and global features, leading to more enriched feature representations and better learning outcomes.
- The channel merging module employs a novel fusion block to extract highly discriminant and domain-relevant features from multiple channel generators. This innovative approach contributes significantly to enhancing the accuracy and performance of the proposed method.
- The proposed technique has shown promising results on benchmark datasets, thereby providing evidence of its potential to be implemented as an effective diagnostic tool for lymphocyte assessment in histology images.

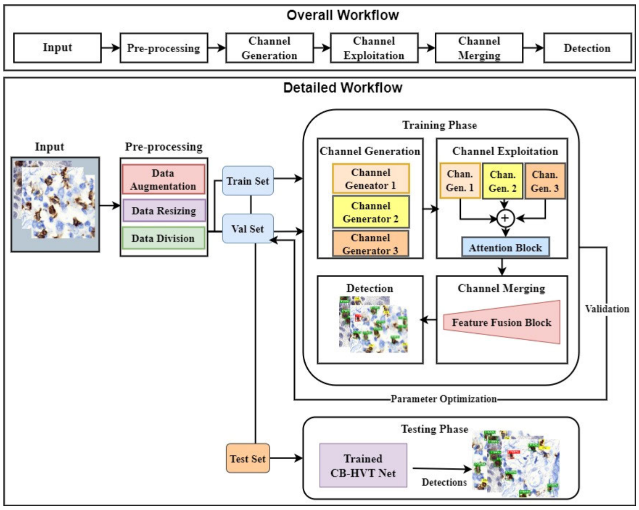

FIGURE 2. **Overall and detailed workflow of the proposed CB-HVT Net. The training and testing phases of the proposed approach are also elaborated.**

The remaining contents are arranged as follows: Related works are briefly summarized in Section II. Section III presents the methodology of the developed CB-HVT Net in detail. We present the experimental results and discuss their implications in Section IV. In Section V we discuss the main findings and finally, Section VI concludes the paper.

II. RELATED WORK

In medical diagnosis, accurate object detection is of utmost importance for precise diagnosis and treatment. Various deep learning architectures have been developed for this purpose, with CNNs and transformers emerging as significant methods for object detection. Taking advantage of their ability to extract features from images and recognize patterns, CNNs, and transformers have been successfully used in medical imaging. Despite their different approaches to image data handling, both CNNs and transformers have shown great potential in medical image analysis.

A. CNN-BASED METHODS

A considerable amount of research has been done using traditional image processing and machine learning-based algorithms for cell identification and detection in digital histopathology slides [28], [29] including region growth [30], morphological operations [31], [32] and hand-crafted feature analysis [33], [34]. However, recent developments in deep learning have revolutionized the field of computer vision, enabling algorithms to learn complex representations from raw data [35]. As a result, deep learning algorithms have been applied to medical image analysis and have demonstrated the ability to perform at levels comparable to and, in some cases, surpassing those of human experts [12], [36], [37], [38]

Janowczyk et al. carried out a study to perform lymphocyte detection in histopathological images of cancer patients. In their work, they proposed a deep learning-based technique to classify lymphocytic and non-lymphocytic patches [39]. Rijthoven et al. carried out lymphocyte detection in breast, colon, and prostate immunohistochemistry cancer images using a YOLO v3-based architecture [40]. Linder et al. proposed a two-stage classification strategy to discriminate lymphocytic WSIs. The first step involved the rough identification of a lymphocytic area, which was then subjected to the second stage to obtain more accurate cell detection results [41]. Swiderska-Chadaj et al. conducted a study to locate lymphocytes in IHC-stained images [42]. In their work, they analyzed the performance of several deep learning models in different lymphocyte containing regions, including artifact, regular, and clustered regions.

Region-based CNNs [43], [44] have shown great performance in detecting complex objects, including lymphocytes [45], [46]. These models exploit a smaller network as a region proposal network (RPN) to identify the probable regions that may contain the object [47]. These selected regions are subsequently analyzed to determine the precise location and nature of the object. Zhang et al. exploited Mask RCNN to develop a unified framework for panoptic segmentation in histology images [48]. Liu et al. carried out cell instance segmentation by incorporating a new module in the segmentation head of Mask RCNN [49]. They also proposed a feature map combination method to integrate the local and global level feature learning. Their findings from experiments demonstrated that integrating this additional module into the model enhanced the precision of cell instance segmentation. Kutlu et al. introduced a computer-aided automated approach that quickly identified and detected different types of WBC in blood images [50]. Zafar et al. proposed a two-phase approach to identify tumor-infiltrating lymphocytes (TILs) in multiple cancer images [51]. Zhang et al. developed a novel architecture to evaluate TILs in hematoxylin and eosin-stained images of breast cancer [52]. Rauf et al. carried out lymphocyte detection using deep CNN-based technique. Their method exploited two different architectures to effectively analyze lymphocytes both at cellular and tissue level in histopathological images [9].

B. TRANSFORMER-BASED METHODS

Transformers, on the other hand, are a newer architecture and have shown significant contributions in object detection tasks specifically in medical imaging [53], [54]. Obeid et al. performed nucleus detection in histopathological images using a transformer [55]. Chen et al. employed transformers for the analysis of gastric histopathological images [56]. They developed two CNN-based modules, named Global Information Module (GIM) and Local Information Module (LIM) for feature extraction. Additionally, they incorporated the Inception-V3 architecture to acquire multi-scale local representations. Although ViT-based methods tackle the limitations of CNN, poor inductive bias and high computation make them unsuitable for many real-time problems, including medical diagnosis [57], [58]. Recently, researchers have come up with the idea of merging both CNNs and transformers to get the benefits of both methods [59], [60],

[61]. In this regard, Srinivas et al., introduced BoTNet where they modified the last three blocks of ResNet and significantly modified the ViT's self-attention mechanism [62]. Guo et al. introduced a transformer architecture in which they added pointwise and depthwise convolution before the self-attention module [63]. Similarly, Chen et al, proposed the first hybrid transformer, TransUNet by combining transformer and UNet for medical image segmentation [20]. Cao et al. introduced Swin-UNet in their work, in which they exploited the Swin attention module for medical image segmentation [64]. Gao et al. utilized UNet architecture for medical image segmentation but replaced the last convolution block with the transformer block to incorporate the attention mechanism [65].

Despite their effectiveness, the above-described methods have their limitations. A major concern is their high computational complexity, which may make them impractical for medical diagnosis in laboratories. Furthermore, some of these methods might not be able to fully capture the relevant information at each stage of the architecture, which could affect their overall performance.

III. METHOD

Automated assessment of lymphocytes in histology images is challenging due to the complex nature of tissue representation. Such complexity often leads to a high percentage of false positives, as well as difficulties in detecting lymphocytes that appear in clusters or exhibit different morphologies [66]. To address these challenges, we have developed a novel framework for lymphocyte assessment. The detailed workflow of the proposed CB-HVT Net is depicted in Figure 3. CB-HVT Net consists of five main modules, named: a) channel generation module, b) channel exploitation module, c) channel merging module, d) region-aware module, and e) classification and detection head. Details of each module are elaborated upon in the following sections.

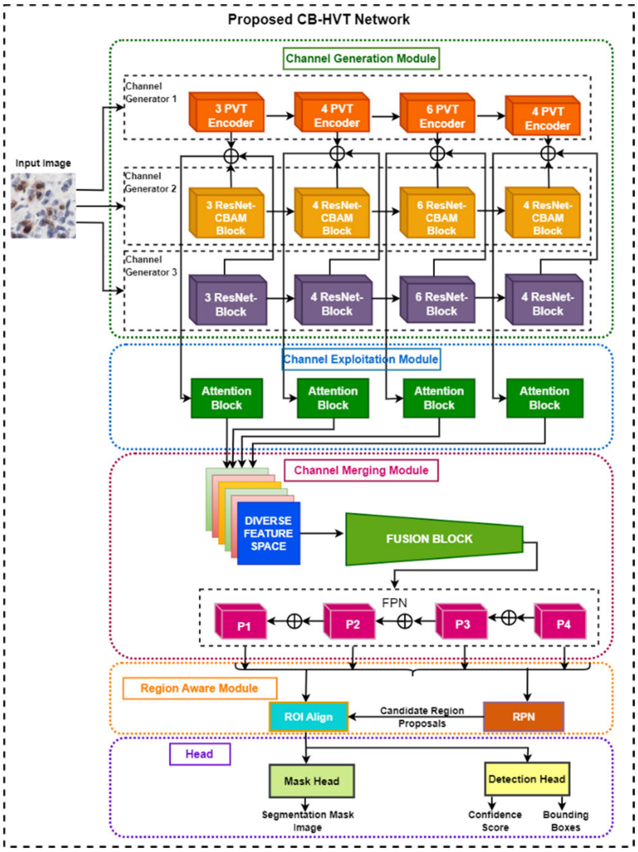

FIGURE 3. Detailed workflow of the proposed architecture CB-HVT Net which has five main modules. In CB-HVT Net three backbone architectures are used which are PVT, ResNet-CBAM, and ResNet. These backbone architectures extract useful features from the images which are then passed to attention blocks in the channel exploitation module where weights are assigned to the features and important features based on higher weight are given greater priority. Channel merging module receivers a dense feature space which is passed through the fusion block, and it reduces the size of the feature space and only gives the most relevant features as output.

A. PROPOSED CHANNEL BOOSTED HYBRID VISION TRANSFORMER NETWORK "CB-HVT NET"

The proposed CB-HVT Net begins by employing its channel generation module to generate boosted channels for a given input image, utilizing transfer learning to produce high-dimensional features. These boosted channels are then passed to the channel exploitation and channel merging modules for channel fusion and reduction, where domain-relevant and discriminant features are re-weighted to enhance their contributions. In the region-aware module, CB-HVT Net uses a Region Proposal Network (RPN) to extract objects containing probable regions. Finally, the classification and detection head generate the final output. Details of each module are described below.

1) CHANNEL GENERATION MODULE

Given the complex patterns of medical images and high-level pattern variations at both tissue and cellular levels, we utilized the idea of channel boosting to generate diverse boosted channels. The proposed CB-HVT Net employs three heterogeneous architectures based on the concepts of vision transformer, spatial and channel attention, and residual connection to capture multi-level variations. The learned multi-variate feature maps from each architecture are added to attain a boosted feature space that not only captures the global level context but also the local image representations in the images (Eq. 1).

$$B_{FS} = \sum_{i=1}^{N} F_{extractor}(I_{MxN}) \tag{1}$$

where in Eq. 1, I_{MxN} is the input image taken by each feature extractor (represented as $F_{extractor}$), and B_{FS} is the generated boosted feature space.

Our channel boosting approach involves domain adaptation-based transfer learning to extract channels from diverse architectures based on their unique learning abilities. To achieve this, we employed two pre-trained CNN-based networks and a transformer-based network.

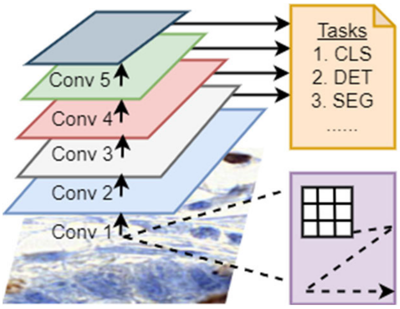

FIGURE 4. Shrinking pyramid structure of PVT model. From the input image, PVT extracts maximum amount of information and keeps on reducing it, based on attention mechanism of transformers, while moving towards the top of pyramid.

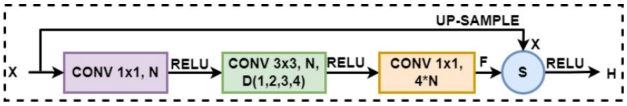

FIGURE 5. Residual network-based channel generator which employs the idea of residual learning to incorporate reference-based learning.

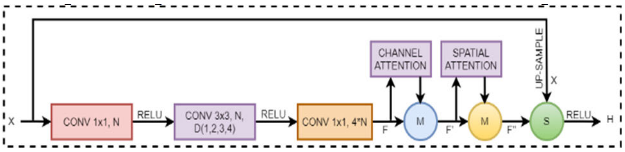

FIGURE 6. Attention-based channel generator which employs the idea of attention with residual connection.

For the transformer-based learner, we utilized the Pyramid Vision Transformer (PVT) to learn long-range dependencies and global-level contextual information [67]. In PVT, a group of convolutional layers initially processes the input image to extract low-level features. Then, the multi-scale transformer module processes these features to extract relevant information at different scales. The output of the transformer is then passed through a set of fully connected layers to produce the final output. The shrinking pyramid-like structure of PVT is shown in Figure 4.

The second channel generator in the CB-HVT Net is a pre-trained 50-layered Residual Network (ResNet-50), which has demonstrated excellent performance in various medical image tasks (Figure 5) [68]. The main idea behind the residual connection is to enable the network to obtain the residual relationship between the input and output of a layer instead of directly learning the underlying mapping (Eq. 2). It utilizes residual connections to enable reference-based learning and solve the problem of dead neurons.

$$Output = Input + F(Input) \tag{2}$$

In Eq. 2, the input is denoted by "*Input*", while "*F*" represents the residual function. The output of the residual block is computed by adding the input and the output of the residual function, which helps to create a shortcut connection between the input and output.

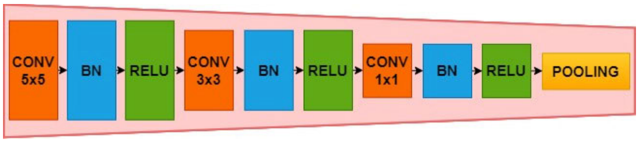

FIGURE 7. Feature merging block used to merge the output of channel generators. The idea of bottleneck is employed in this merging block to obtain the most relevant features as output.

Moreover, to capture class-specific features at both the channel and spatial levels, attention-based ResNet is employed. The architecture of the attention-based Channel Generator is shown in Figure 6. Eqs. 3 and 4 illustrate the concept of spatial and channel attention, respectively.

$$F' = M_C\,(F) \otimes F) \tag{3}$$

$$F'' = M_S\left(F'\right) \otimes F') \tag{4}$$

here, the symbol ⊗ denotes element-wise multiplication, while M_s(F) and M_c(F) represent spatial attention and channel attention, respectively. The input feature map is represented by F. To refine the feature map, we perform an element-wise multiplication between the channel attention and the input feature map, resulting in a refined feature map F'. This refined feature map is then used in another element-wise multiplication with the spatial attention, resulting in the refined output feature map F''. This process helps to enhance the feature map representation and improve the model's performance.

The utilization of these three diverse channel generators in HVT-CB Net resulted in the creation of a diverse and boosted feature space, which improves the ability to discriminate and identify objects with distinct boundaries, variable sizes, and shapes.

2) CHANNEL EXPLOITATION MODULE

The learned diverse and boosted channels from the channel generation module are exploited in the channel exploitation module to figure out the domain-relevant channels (Eq. 5). The boosted feature maps from diverse learners are analyzed using the attention mechanism (Eqs. 3 and 4) to allow the network to focus on the most relevant channels while ignoring the ones with redundant information.

$$R_{FS} = CE_{module}(B_{FS}) \tag{5}$$

In Eq. 5, CE_{module} is the channel exploitation module that takes B_{FS}) as input and generates a refined feature space, expressed by R_{FS}.

3) CHANNEL MERGING MODULE

The proposed CB-HVT Net employs a novel feature merging block to systematically reduce the channel dimension while retaining the most relevant features from the aggregated feature space. This block improves the accuracy and representation capacity of the proposed approach. Figure 7 illustrates the design of this feature merging block, which applies several sets of transformations, such as 3 × 3 and 1 × 1 convolutions, to the outputs of various channel generators. Additionally, a feature pyramid network is also utilized in this module to extract feature maps at multiple levels of abstraction. This enables the model to capture high-level morphological and textural variations in lymphocytes. The channel merging module is expressed mathematically in Eq. 6.

$$M_{FS} = FPN(FB\left(R_{FS}\right)) \tag{6}$$

where, FB is the fusion block, and M_{FS} is the merged feature space, output of the channel merger module.

4) REGION AWARE MODULE

The region-aware module of the proposed CB-HVT employs a Region Proposal Network (RPN) to identify probable regions that may contain lymphocytes. The selected region proposals are then passed to the ROI Align layer, which resizes all the feature maps to align them with their respective proposal region. It utilizes bilinear interpolation for the accurate sampling of fixed-size feature maps without compromising the spatial resolution of the original feature maps.

5) DETECTION AND SEGMENTATION HEAD

These fixed-sized feature maps are fed to the detection and segmentation head to detect and localize the lymphocytes [69]. The detection head produces a set of bounding boxes for possible lymphocytic objects and their corresponding objectness scores, whereas the segmentation head produces a binary mask for each lymphocyte, indicating its precise location in the image.

B. ERROR FUNCTION OF THE PROPOSED CB-HVT NET

The error function of the proposed approach is a combined loss for each of its detection and segmentation heads, given in Eq. 7, where L_C is the Cross-Entropy Loss, L_L is L1 Loss, and L_B is the Binary Cross Entropy Loss.

$$L_{CB-HVI} = L_C + L_L + L_B \tag{7}$$

The detection head employs two loss functions, the Cross-Entropy and L1 loss for the prediction of the bounding box and class label, respectively (Eq. 8 and Eq. 9). The segmentation head of the proposed CB-HVT Net is the Binary Cross Entropy loss to predict the binary mask for each object in the image (Eq. 10).

$$L_C = -log(p\_j[y\_j])) \tag{8}$$

$$L_L = \frac{SUM\left|t_j - t*\_j\right|}{N}) \tag{9}$$

$$L_B = \frac{-1}{N} * sum\left(y_j * \log\left(p_j\right) + \left(1 - y_j\right) * \log\left(1 - p_j\right)\right) \tag{10}$$

The cross-entropy loss and L1 loss are represented in Eqs. (8 & 9), p_j in Eq. 8 is the likelihood. j is the anchor and y is the actual and predicted class labels. In Eq. 9 t_j is the actual bounding box coordinates of the j_{th} anchor while $t*\_j$ is the predicted bounding box coordinate. In Eq.10, y_j is the true label and p_j is the probability of the predicted class.

TABLE 1. Details of LYSTO and NuClick datasets are presented below.

Dataset details	LYSTO dataset	NuClick dataset
Image size	267 x 267 pixels	256x256 pixels
Staining	Immunohistochemistry	Immunohistochemistry
Train set	9000	471
Val. set	3000	99
Test set	3000	300

C. COMPARATIVE STUDY WITH EXISTING METHODS

For a fair comparison, we compared our proposed CB-HVT Net with other existing models for lymphocyte assessment. In this regard, we selected MaskRCNN as the state-of-the-art two-stage detector, YOLO and SC-Net as the latest single-stage detectors, and Unet as a semantic segmentation model. We evaluated our proposed CB-HVT Net against these existing models on the test sets of the LYSTO and NuClick datasets to determine its effectiveness.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASETS

In this study, we utilized two datasets for the training of the proposed method and three datasets for the testing of the proposed "CB-HVT Net". Training sets included image samples from the LYSTO and the NuClick datasets, and for testing we utilized test sets of LYSTO, NuClick and LYON datasets. The LYSTO dataset was released by the Lymphocyte Assessment and Hackathon (LYSTO) challenge organizers [70]. It consists of 43 patients for breast colon and prostate cancer and released a total of 20k images. Among these, we selected 19 patients for training, 9 patients for validation, and 6 patients for testing. In contrast, the NuClick dataset was introduced by Koohbanani et al. and contained 871 images from 440 WSIs [71]. Keeping in consideration, the slide number we selected 471 images for training, 99 for validation, and 300 for testing of our model. The datasets were divided based on the WSIs count for different types of cancers. The training and testing sets were prepared to mimic the real-time diagnosis and to avoid the tissue slide mixing across the patient ID. Details of these datasets are depicted in Table 1.

B. TRAINING AND IMPLEMENTATION DETAILS

All the experiments of this work were conducted on an NVIDIA GTX machine with 1070 GPU and 8GB of memory, using an open-source PyTorch. Moreover, an open-source toolbox OpenMMLab was employed for the implementation of all the proposed and comparative models. The rest of the libraries and their versions are listed in Table 2, whereas

TABLE 2. The details of libraries and their version.

Libraries	Versions
Pytorch	1.12.1
Numpy	1.23.1
Opencv-python	4.6.0.66
CUDA version	11.8

TABLE 3. Details of the selected hyperparameters for the proposed and comparative models.

Configuration	Value
Epochs	30
Learning Rate	0.0025
Weight Decay	0.0001
Momentum	0.9
Optimizer	SGD
Batch size	4

Table 3 lists the hyper-parameters for all the models. These hyper-parameters were optimized using the validation set.

C. PERFORMANCE METRICS

F-Score and recall were used as the performance metrics to analyze the performance of the proposed model and the comparison models. The F-Score, which is the harmonic mean of precision and recall, is an unbiased and reliable measure, to evaluate the representation learning ability of the model especially when the data is imbalanced. The model's Recall indicates its ability to reliably identify true predictions. The formulas for F-Score and recall are shown in Eq. 11 and Eq. 12, respectively.

$$F - score = \frac{2\,(Precision * Recall)}{Precision + Recall} \tag{11}$$

$$Recall = \frac{TruePositive}{(True\ Positive + False\ Negative)} \tag{12}$$

D. QUANTITATIVE RESULTS OF THE PROPOSED CB-HVT NET COMPARED TO THE EXISTING ARCHITECTURES

The quantitative results of the proposed CB-HVT Net and the comparative models are shown in Table 4. These models are evaluated based on F-Score and recall. The proposed CB-HVT Net outperformed Mask RCNN, SC-Net, and YOLO with an F-Score of 0.88 and 0.82 on LYSTO and NuClick datasets, respectively. In terms of recall, SC-Net performed well on the LYSTO dataset as compared to Mask RCNN and YOLO, whereas YOLO showed an improved recall when evaluated on the NuClick dataset. The results are presented in Table 4.

E. QUALITATIVE RESULTS OF THE PROPOSED CB-HVT NET

To further understand the effectiveness of the proposed CB-HVT Net, a qualitative analysis was conducted in addition to the quantitative evaluation. The outcomes of our model while evaluating lymphocytes in the LYSTO and NuClick datasets are shown in Figure 8. The findings clearly show that our proposed CB-HVT Net performs better not only in terms of quantitative evaluation but also in terms of visual interpretation.

TABLE 4. Comparison of the proposed CB-HVNet on LYSTO and NuClick with state-of-the-art models.

Technique	LYSTO		NuClick	
	F-Score	Recall	F-Score	Recall
Proposed CB-HVT Net	0.88	0.93	0.82	0.99
Mask-RCNN	0.80	0.73	0.78	0.67
SC-Net	0.85	0.82	0.80	0.84
YOLO	0.80	0.69	0.76	0.86

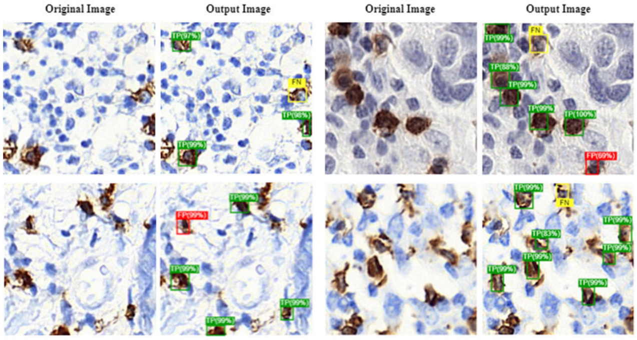

FIGURE 8. Results of the proposed CB-HVT Net on LYSTO and NuClick datasets (Legend: green: TP, red: FP, yellow: FN).

F. ABLATION STUDY

We carried out extensive experiments to assess the efficacy of our proposed architecture design. Specifically, we performed various experiments to identify the optimal combination of channel generators in the channel generation module and the fusion block in the channel merging module. Different configurations of these models are referred to as comparison models 1-6.

1) CHANNEL GENERATION MODULE

This section describes the channel generation module, which consists of several channel generators used to generate boosted channels. We experimented with different combinations of architectures to identify the best-performing model. Specifically, we employed ResNet50, ResNet50 with CBAM, Autoencoder, and PVT backbone, with various combinations. Table 5 summarizes the different architecture combinations used for channel generation.

2) CHANNEL MERGING MODULE

Experiments with different combinations of fusion blocks are done to identify the most optimal method for merging the generated diverse boosted channels. Table 6 presents the architecture of these channel mergers, and Table 7 reports their results in terms of F-Score and recall. The results presented in Table 7 are all self-implemented models that have different channel generation and merging blocks and were used to compare the results with the proposed CB-HVT Net. Results show that the combination of Channel Merger-6 with Channel Generator-6 performed better than the other

TABLE 5. Architectural details for different channel generators.

Model	Description
Channel Generator-1	ResNet, PVT
Channel Generator-2	ResNet-50, ResNet-CBAM, PVT, Convolutional Autoencoder
Channel Generator-3	ResNet-50, ResNet-CBAM, ConvAutoencoder
Channel Generator-4	ResNet-50, ResNet-CBAM, PVT, ResNet-101
Channel Generator-5	ResNet-50, ResNet-CBAM, ResNext, ResNet-101
Channel Generator-6	ResNet-CBAM, ResNext

TABLE 6. Architectural details of several fusion blocks utilized in the feature merger module.

Model	Fusion Block Architectures
Channel Merger-1	5x5 conv, Batch Norm, ReLU, 3x3 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling
Channel Merger-2	5x5 conv, Batch Norm, ReLU, 3x3 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling
Channel Merger-3	5x5 conv, Batch Norm, ReLU, 3x3 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling
Channel Merger-4	7x7 conv, Batch Norm, ReLU, 5x5 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling
Channel Merger-5	3x3 conv, Batch Norm, ReLU, 3x3 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling
Channel Merger-6	5x5 conv, Batch Norm, ReLU, 3x3 conv, Batch Norm, ReLU, 1x1 conv, Batch Norm, ReLU, Pooling

TABLE 7. Comparison of various settings for the proposed "CB-HVT Net."

Backbone	LYSTO		NuClick	
	F-Score	Recall	F-Score	Recall
Comparison Model-1 (Ch. Gen-1 + Ch. M-1)	86.60	92.12	85.00	99.78
Comparison Model-2 (Ch. Gen-2 + Ch. M-2)	86.54	91.35	83.48	99.89
Comparison Model-3 (Ch. Gen-3 + Ch. M-3)	82.54	89.25	81.72	98.05
Comparison Model-4 (Ch. Gen-4 + Ch. M-4)	85.45	90.11	80.65	99.43
Comparison Model-5 (Ch. Gen-5 + Ch. M-5)	87.54	91.12	80.49	99.64
Comparison Model-6 (Ch. Gen-6 + Ch. M-6)	88.19	93.56	82.14	99.64

combinations, but the proposed CB-HVT Net performed better in comparison with all six in-house built comparison architectures.

3) GENERALIZATION ANALYSIS

We also evaluated the performance of our proposed CB-HVT Net on an unseen test set, which was released as part of the LYON'19 challenge. We tested the performance of our proposed model on the LYON dataset as a supplementary study as the challenge organizers did not provide any training set for the test images. However, we were able to evaluate our pre-trained model on the provided 441 ROI images. While the test labels were kept hidden by the challenge organizers, we obtained the results from the leaderboard. The results indicate that the proposed CB-HVT Net achieved an F-Score of 0.70 and a recall of 0.64 in identifying lymphocytes in these images (Figures 7 and 8).

V. DISCUSSION

In this work, we have developed a deep learning-based approach "CB-HVT Net" to identify lymphocytes in multi-cancer histopathological images. Automated lymphocyte detection can assist physicians and reduce their burden while counting lymphocytes manually and performing disease diagnosis.

The proposed CB-HVT Net is developed as a hybrid model by utilizing both CNNs and ViTs as feature extractors. Exploitation of the idea of channel boosting in the proposed approach helped in dealing with the heterogeneous morphology of lymphocytes.

This study also demonstrates that by combining the learning capabilities of diverse architectures we can achieve a boosted and diverse feature space that can deal with the diversity at both tissue and cellular levels. In addition, utilization of the proposed novel feature fusion block retained the most relevant features by re-weighting the feature maps using the spatial and channel attention mechanisms.

Deep learning models require large amounts of training data and medical images often face the issue of the unavailability of pathologist labeled datasets. In this work, we utilized domain adaptation-based transfer learning by fine-tuning the pre-trained feature extractors. This approach implies that deep models trained on other domains can be transferred to medical images with some fine tuning.

The study further demonstrates that by employing FPN and RPN lymphocytes can be effectively captured even in the presence of regions with artifacts and the issue of clustered lymphocytes can be tackled. The generalization capacity of the proposed CB-HVT Net was evaluated on an unseen test set where it demonstrated notable performance. Additionally, the better performance of the developed model as compared to other existing architectures shows its significance as an assistance tool for pathologists. The proposed CB-HVT Net still has some margin for improvement. In the future, more optimal configurations of the feature extractors and the merging strategy can be found by extensive experimentation with different deep fusion block combinations. Such efforts would enhance the performance of the model and increase its applicability in medical image analysis.

VI. CONCLUSION

Accurate and automated lymphocyte assessment plays an important role in cancer analysis, offering significant prognostic value. Due to varied appearance of lymphocytes, unavailability of pathologist labeled datasets, presence of artifacts and the clustered presence of lymphocytes, its detection is very challenging. In this work, a Channel Boosted

Hybrid Vision Transformer ''CB-HVT'' is proposed to identify lymphocytes in IHC-stained histology images. In the proposed approach, channel generation, channel exploitation, and channel merging modules are utilized to increase the model's representation learning ability. The channel generator module of the proposed CB-HVT Net exploits both CNN-based and transformer-based architectures to generate informative channels, enabling robust feature extraction. These channels are then systematically merged and fused using a novel channel merging branch, facilitating the integration of the most domain-relevant feature maps.

The proposed CB-HVT Net detected lymphocytes even in the existence of significant intra-class resemblance, occlusion, and artifacts due to the flaws in the lens and blurring effects. The proposed technique also employed a region-aware module to initially identify the probable lymphocytic regions, which were later processed by the detection and segmentation head to detect lymphocytes. We evaluated the performance of CB-HVT Net by comparing it with various state-of-the-art architectures. The results indicate that the proposed approach outperforms other existing models in terms of F-Score and recall. The superior performance of the CB-HVT Net is attributed to its effective channel generation and merging techniques, as well as its utilization of channel boosting and transfer learning. The proposed CB-HVT Net holds great potential for enhancing the efficacy and accuracy of cancer diagnosis and treatment planning by providing pathologists with an assistance tool that can act as a second opinion.

ACKNOWLEDGMENT

This research was supported by ''Regional Innovation Strategy (RIS)'' through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-001).

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, ''Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,'' *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] B. S. Chhikara and K. Parang, ''Global Cancer Statistics 2022: The trends projection analysis,'' *Chem. Biol. Lett.*, vol. 10, no. 1, p. 451, 2023.

[3] S. Benavente, A. Sánchez-García, S. Naches, M. E. LLeonart, and J. Lorente, ''Therapy-induced modulation of the tumor microenvironment: New opportunities for cancer therapies,'' *Frontiers Oncol.*, vol. 10, Oct. 2020, Art. no. 582884, doi: 10.3389/FONC.2020.582884/BIBTEX.

[4] R. Salgado, ''The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: Recommendations by an international TILs working group 2014,'' *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, Feb. 2015, doi: 10.1093/ANNONC/MDU450.

[5] E. Garcia, R. Hermoza, C. B. Castanon, L. Cano, M. Castillo, and C. Castanneda, ''Automatic lymphocyte detection on gastric cancer IHC images using deep learning,'' in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 200–204, doi: 10.1109/CBMS.2017.94.

[6] J. Chen and C. Srinivas, ''Automatic lymphocyte detection in H&E images with deep neural networks,'' 2016, *arXiv:1612.03217*.

[7] S. W. Jahn, M. Plass, and F. Moinfar, ''Digital pathology: Advantages, limitations and emerging perspectives,'' *J. Clin. Med.*, vol. 9, no. 11, p. 3697, Nov. 2020, doi: 10.3390/JCM9113697.

[8] N. A. Barsha, A. Rahman, and M. R. C. Mahdy, ''Automated detection and grading of invasive ductal carcinoma breast cancer using ensemble of deep learning models,'' *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104931, doi: 10.1016/J.COMPBIOMED.2021.104931.

[9] Z. Rauf, A. Sohail, S. H. Khan, A. Khan, J. Gwak, and M. Maqbool, ''Attention-guided multi-scale deep object detection framework for lymphocyte analysis in IHC histological images,'' *Microscopy*, vol. 72, no. 1, pp. 27–42, Feb. 2023, doi: 10.1093/jmicro/dfac051.

[10] J. Gao, Q. Jiang, B. Zhou, and D. Chen, ''Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview,'' *Math. Biosci. Eng.*, vol. 16, no. 6, pp. 6536–6561, 2019, doi: 10.3934/MBE.2019326.

[11] A. Sohail, A. Khan, H. Nisar, S. Tabassum, and A. Zameer, ''Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier,'' *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102121.

[12] S. Graham, ''Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images,'' *Med. Image Anal.*, vol. 58, pp. 1–18, Dec. 2019, doi: 10.1016/j.media.2019.101563.

[13] S. H. Khan, A. Sohail, A. Khan, M. Hassan, Y. S. Lee, J. Alam, A. Basit, and S. Zubair, ''COVID-19 detection in chest X-ray images using deep boosted hybrid learning,'' *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104816, doi: 10.1016/J.COMPBIOMED.2021.104816.

[14] A. Ben Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, and C. Wemmert, ''Deep learning for colon cancer histopathological images analysis,'' *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104730, doi: 10.1016/J.COMPBIOMED.2021.104730.

[15] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, ''A survey of the recent architectures of deep convolutional neural networks,'' *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/s10462-020-09825-6.

[16] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, ''Toward transformer-based object detection,'' 2020, *arXiv:2012.09958*.

[17] I. Khalfaoui-Hassani, T. Pellegrini, and T. Masquelier, ''Dilated convolution with learnable spacings,'' 2021, *arXiv:2112.03740*.

[18] L. Yang, Q. Song, Y. Wu, and M. Hu, ''Attention inspiring receptive-fields network for learning invariant representations,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1744–1755, Jun. 2019, doi: 10.1109/TNNLS.2018.2873722.

[19] T. Dhamija, A. Gupta, S. Gupta, R. Katarya, and G. Singh, ''Semantic segmentation in medical images through transfused convolution and transformer networks,'' *Appl. Intell.*, vol. 53, no. 1, pp. 1132–1148, Jan. 2023, doi: 10.1007/S10489-022-03642-W/FIGURES/9.

[20] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, ''TransUNet: Transformers make strong encoders for medical image segmentation,'' 2021, *arXiv:2102.04306*.

[21] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, ''HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6191–6201, doi: 10.1109/WACV56688.2023.00614.

[22] H. Wu, ''CvT: Introducing convolutions to vision transformers,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2021, pp. 22–31.

[23] A. Güngör, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, ''TranSMS: Transformers for super-resolution calibration in magnetic particle imaging,'' *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3562–3574, Dec. 2022, doi: 10.1109/TMI.2022.3189693.

[24] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, ''ViTAE: Vision transformer advanced by exploring intrinsic inductive bias,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 28522–28535.

[25] O. Moutik, ''Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?'' *Sensors*, vol. 23, no. 2, p. 734, Jan. 2023, doi: 10.3390/S23020734.

[26] C. Li, T. Tang, G. Wang, J. Peng, B. Wang, X. Liang, and X. Chang, ''BossNAS: Exploring hybrid CNN-transformers with block-wisely self-supervised neural architecture search,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12261–12271.

[27] S. Pan, X. Liu, N. Xie, and Y. Chong, ''EG-TransUNet: A transformer-based U-Net with enhanced and guided models for biomedical image segmentation,'' *BMC Bioinf.*, vol. 24, no. 1, p. 85, Mar. 2023, doi: 10.1186/s12859-023-05196-1.

[28] S. Cohen, ''The evolution of machine learning: Past, present, and future,'' in *Artificial Intelligence and Deep Learning in Pathology*. Elsevier, Jan. 2021, pp. 1–12, doi: 10.1016/B978-0-323-67538-3.00001-4.

[29] S. Cagnoni, E. Lutton, and G. Olague, *Genetic and Evolutionary Computation for Image Processing and Analysis*. vol. 8. New York, NY, USA: Hindawi Publishing Corporation, 2008.

[30] Z. Xu, C. F. Moro, D. Kuznyecov, B. Bozóky, L. Dong, and Q. Zhang, ''Tissue region growing for hispathology image segmentation,'' in *Proc. 3rd Int. Conf. Biomed. Imag., Signal Process.*, Oct. 2018, pp. 86–92, doi: 10.1145/3288200.3288213.

[31] J. M. R. Mancha, V. Meas-Yedid, S. V. Martínez, J. C. Olivo-Marin, and G. Stamon, ''Morphological active contours for image segmentation,''

[32] S. Chatterjee, D. Dey, and S. Munshi, ''Integration of morphological preprocessing and fractal based feature extraction with recursive feature elimination for skin lesion types classification,'' *Comput. Methods Programs Biomed.*, vol. 178, pp. 201–218, Sep. 2019, doi: 10.1016/J.CMPB.2019.06.018.

[33] L. C. Faria, L. F. Rodrigues, J. F. Mari, L. C. de Faria, and J. F. Mari, ''Cell classification using handcrafted features and bag of visual words,'' presented at the XIV Workshop de Visão Computacional (WVC), 2018. [Online]. Available: http://homes.di.unimi.it/scotti/all/

[34] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, ''Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features,'' *J. Med. Imag.*, vol. 1, no. 3, Oct. 2014, Art. no. 034003, doi: 10.1117/1.JMI.1.3.034003.

[35] Y. Li, B. Sixou, and F. Peyrin, ''A review of the deep learning methods for medical images super resolution problems,'' *IRBM*, vol. 42, no. 2, pp. 120–133, Apr. 2021, doi: 10.1016/J.IRBM.2020.08.004.

[36] S. Graham, ''MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images,'' *Med. Image Anal.*, vol. 52, pp. 199–211, Feb. 2019, doi: 10.1016/j.media.2018.12.001.

[37] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, ''Deep learning in medical image analysis,'' in *Deep Learning in Medical Image Analysis*, 2020, pp. 3–21.

[38] Y. Cui, G. Zhang, Z. Liu, Z. Xiong, and J. Hu, ''A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images,'' *Med. Biol. Eng. Comput.*, vol. 57, no. 9, pp. 2027–2043, Sep. 2019, doi: 10.1007/s11517-019-02008-8.

[39] A. Janowczyk and A. Madabhushi, ''Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,'' *J. Pathol. Informat.*, vol. 7, no. 1, p. 29, Jan. 2016, doi: 10.4103/2153-3539.186902.

[40] M. V. Rijthoven, Z. Swiderska-Chadaj, K. Seeliger, J. V. D. Laak, and F. Ciompi, ''You only look on lymphocytes once,'' Tech. Rep., 2018.

[41] N. Linder, J. C. Taylor, R. Colling, R. Pell, E. Alveyn, J. Joseph, A. Protheroe, M. Lundin, J. Lundin, and C. Verrill, ''Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours,'' *J. Clin. Pathol.*, vol. 72, no. 2, pp. 157–164, Feb. 2019, doi: 10.1136/jclinpath-2018-205328.

[42] Z. Swiderska-Chadaj, H. Pinckaers, M. van Rijthoven, M. Balkenhol, M. Melnikova, O. Geessink, Q. Manson, M. Sherman, A. Polonia, J. Parry, M. Abubakar, G. Litjens, J. van der Laak, and F. Ciompi, ''Learning to detect lymphocytes in immunohistochemistry with deep learning,'' *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101547, doi: 10.1016/j.media.2019.101547.

[43] R. Xue, W. Xiang, and Y. Deng, ''Improved faster R-CNN based on CSP-DPN,'' *Proc. Comput. Sci.*, vol. 199, pp. 1490–1497, Jan. 2022, doi: 10.1016/j.procs.2022.01.190.

[44] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, ''CLU-CNNs: Object detection for medical images,'' *Neurocomputing*, vol. 350, pp. 53–59, Jul. 2019, doi: 10.1016/j.neucom.2019.04.028.

[45] J. W. Johnson, ''Automatic nucleus segmentation with mask-RCNN,'' in *Proc. Sci. Inf. Conf.*, vol. 944, Apr. 2019, pp. 399–407, doi: 10.1007/978-3-030-17798-0_32.

[46] I. K. Evangeline, J. G. Precious, N. Pazhanivel, and S. P. A. Kirubha, ''Automatic detection and counting of lymphocytes from immunohistochemistry cancer images using deep learning,'' *J. Med. Biol. Eng.*, vol. 40, no. 5, pp. 735–747, Oct. 2020, doi: 10.1007/s40846-020-00545-4.

[47] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[48] D. Zhang, Y. Song, D. Liu, H. Jia, S. Liu, Y. Xia, H. Huang, and W. Cai, ''Panoptic segmentation with an end-to-end cell R-CNN for pathology image analysis,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 11071, pp. 237–244, 2018, doi: 10.1007/978-3-030-00934-2_27/COVER.

[49] D. Liu, D. Zhang, Y. Song, C. Zhang, F. Zhang, L. O'Donnell, and W. Cai, ''Nuclei segmentation via a deep panoptic model with semantic feature fusion,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 861–868, doi: 10.24963/IJCAI.2019/121.

[50] H. Kutlu, E. Avci, and F. Özyurt, ''White blood cells detection and classification based on regional convolutional neural networks,'' *Med. Hypotheses*, vol. 135, Feb. 2020, Art. no. 109472, doi: 10.1016/J.MEHY.2019.109472.

[51] M. M. Zafar, Z. Rauf, A. Sohail, A. R. Khan, M. Obaidullah, S. H. Khan, Y. S. Lee, and A. Khan, ''Detection of tumour infiltrating lymphocytes in CD3 and CD8 stained histopathological images using a two-phase deep CNN,'' *Photodiagnosis Photodynamic Therapy*, vol. 37, Mar. 2022, Art. no. 102676, doi: 10.1016/j.pdpdt.2021.102676.

[52] X. Zhang, X. Zhu, K. Tang, Y. Zhao, Z. Lu, and Q. Feng, ''DDT-Net: A dense dual-task network for tumor-infiltrating lymphocyte detection and segmentation in histopathological images of breast cancer,'' *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102415, doi: 10.1016/j.media.2022.102415.

[53] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, ''Transformers in medical imaging: A survey,'' *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802, doi: 10.1016/j.media.2023.102802.

[54] A. R. Khan and A. Khan, ''MaxViT-UNet: Multi-axis attention for medical image segmentation,'' 2023, *arXiv:2305.08396*.

[55] A. Obeid, T. Mahbub, S. Javed, J. Dias, and N. Werghi, ''NucDETR: End-to-end transformer for nucleus detection in histopathology images,'' in *Proc. Int. Workshop Comput. Math. Modeling Cancer Anal.*, vol. 13574, 2022, pp. 47–57, doi: 10.1007/978-3-031-17266-3_5/COVER.

[56] H. Chen, C. Li, G. Wang, X. Li, M. Mamunur Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, S. Ai, and M. Grzegorzek, ''GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection,'' *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108827, doi: 10.1016/J.PATCOG.2022.108827.

[57] L. Morra, L. Piano, F. Lamberti, and T. Tommasi, ''Bridging the gap between natural and medical images through deep colorization,'' in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 835–842, doi: 10.1109/ICPR48806.2021.9412444.

[58] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, ''ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond,'' *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1141–1162, May 2023, doi: 10.1007/s11263-022-01739-w.

[59] A. Khan, Z. Rauf, A. Sohail, A. Rehman, H. Asif, A. Asif, and U. Farooq, ''A survey of the vision transformers and its CNN-transformer based variants,'' 2023, *arXiv:2305.09880*.

[60] C. Ge, Y. Liang, Y. Song, J. Jiao, J. Wang, and P. Luo, ''Revitalizing CNN attention via transformers in self-supervised visual representation learning,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 4193–4206.

[61] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, ''LeViT: A vision transformer in ConvNet's clothing for faster inference,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12239–12249.

[62] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, ''Bottleneck transformers for visual recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524, doi: 10.1109/CVPR46437.2021.01625.

[63] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, ''CMT: Convolutional neural networks meet vision transformers,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12165–12175, doi: 10.1109/CVPR52688.2022.01186.

[64] H. Cao, ''Swin-UNet: UNet-like pure transformer for medical image segmentation,'' in *Proc. Eur. Conf. Comput. Vis.*, May 2021, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[65] Y. Gao, M. Zhou, and D. N. Metaxas, ''UTNet: A hybrid transformer architecture for medical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 12903, 2021, pp. 61–71, doi: 10.1007/978-3-030-87199-4_6/COVER.

[66] Z. Rauf, A. R. Khan, A. Sohail, H. Alquhayz, J. Gwak, and A. Khan, "Lymphocyte detection for cancer analysis using a novel fusion block based channel boosted CNN," *Sci. Rep.*, vol. 13, no. 1, Aug. 2023, Art. no. 14047, doi: 10.1038/s41598-023-40581-z.

[67] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[68] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[69] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.

[70] Y. Jiao, "LYSTO: The lymphocyte assessment hackathon and benchmark dataset," 2023, *arXiv:2301.06304*.

[71] N. A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot, "NuClick: A deep learning framework for interactive segmentation of microscopic images," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101771, doi: 10.1016/j.media.2020.101771.

MOMINA LIAQAT ALI received the bachelor's and master's degrees in computer science from the Pakistan Institute of Engineering and Applied Sciences (PIEAS). She is interested in finding solutions to real-world problems through computer vision along with deep learning.

ZUNAIRA RAUF received the B.S. degree in bioinformatics and the M.S. degree in computational science and engineering (CS&E). She is currently pursuing the Ph.D. degree with the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS). Her research interests include deep neural networks, biomedical informatics, machine learning, computer vision, and medical image analysis.

ASIFULLAH KHAN is currently a Professor with the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS). He is also leading the Pattern Recognition Laboratory and the PIEAS Artificial Intelligence Center. His research interests include deep neural networks, machine learning, pattern recognition, intrusion detection, medical image analysis, and digital watermarking. He has received the Prestigious "Pride of Performance" Award in computer science from the President of Pakistan.

ANABIA SOHAIL received the B.S. and M.S. degrees in bioinformatics and the Ph.D. degree in computer sciences from the Pakistan Institute of Engineering and Applied Sciences (PIEAS). She is currently a Postdoctoral Researcher in computer sciences with Khalifa University, Abu Dhabi. Her research interests include deep neural networks, machine learning, biomedical informatics, and medical image analysis.

RAFI ULLAH received the M.S. degree in computer system engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, in 2006, and the Ph.D. degree in computer and information sciences from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan, in 2010. From 2011 to 2012, he was a Postdoctoral Fellow with Universiti Teknologi PETRONAS. He is currently an Associate Professor with Universiti Teknologi PETRONAS, Malaysia. Before this, he was an Associate Professor with the National University of Technology, Islamabad, Pakistan, and an Assistant Professor with Majmaah University, Saudi Arabia, and COMSATS Islamabad, Pakistan. He has published several peer-reviewed journals, conference articles, and patents. His research interests include computer vision, machine learning, multimedia security, medical imaging, and brain/EEG signals.

JEONGHWAN GWAK received the Ph.D. degree in machine learning and artificial intelligence from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2014. From 2002 to 2007, he worked for several companies and research institutes as a Researcher and the Chief Technician. From 2014 to 2016, he was a Postdoctoral Researcher with GIST, and from 2016 to 2017, as a Research Professor. From 2017 to 2019, he was a Research Professor with the Biomedical Research Institute and the Department of Radiology, Seoul National University Hospital, Seoul, South Korea. In 2019, he joined the Korea National University of Transportation (KNUT), Chungju, Republic of Korea, as an Assistant Professor, where he has been an Associate Professor, since 2021. He is also the Director of the Algorithmic Machine Intelligence Laboratory. His current research interests include deep learning, computer vision, image and video processing, AIoT, fuzzy sets and systems, evolutionary algorithms, optimization, and relevant applications of medical and visual surveillance systems.