Enhancing Model Generalization of Cervical Fluid-Based Cell Detection through Causal Feature Extraction:A Novel Method

Qiao Pan PANQIAO@DHU.EDU.CN
Bin Yang YANGBIN_DHU@126.COM
Dehua Chen CHENDEHUA@DHU.EDU.CN
Mei Wang WANGMEI@DHU.EDU.CN
School of Computer Science and Technology, Donghua University, Shanghai, China

Editors: Berrin Yanıkoğlu and Wray Buntine

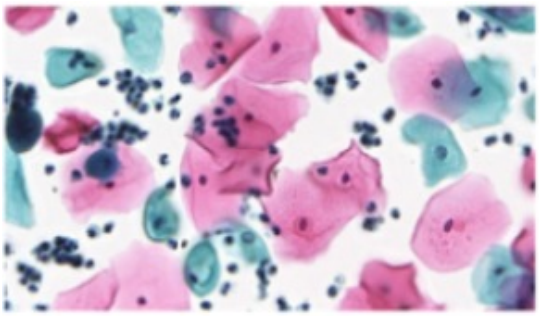
Abstract

Cervical cancer is the most common gynecologic malignancy, and in clinical practice, cervical cancer is best treated if it is detected at an early stage. Thinprep Cytologic Test (TCT) is the best early detection method for cervical cancer as determined by the WHO. As the coverage of early detection of cervical cancer increases, the number of samples in hospitals increases annually, and the pressure on the pathologists to read the cytological images increases, which easily leads to an increase in the rate of misdiagnosis and missed diagnosis. Therefore, automatic detection of abnormal cells in cervical cytology images of cervical fluid using deep learning techniques has become a hot research topic today. However, existing deep learning models for cell detection often collect a single data source from a medical institution for construction. Different medical institutions have different equipment and staining methods, and the accuracy, magnification, and staining results of the images obtained will be different. As a result, the application performance of the model in different medical institution data is not good, and there is a problem of domain shift. To address these problems, this paper proposes a method for cervical fluid-based cell detection based on causal feature extraction. The method is based on the one-stage detection model RetinaNet, and incorporates causal autoencoder to learn the invariant causal feature representation from data. It reduces the impact of task-irrelevant feature representations, reduces the variability of feature distributions in different datasets, and effectively solves the domain shift problem. The addition of deformable convolution and attention mechanism enhances the feature extraction capability for foreground categories with variable shapes in cervical fluid-based pathology images. This reduces the impact of possible strong correlation between background features and goal cells, and reduces the interference of the foreground categories by fading and lack of brightness in the staining. The generalization ability of the model is improved, which makes the model better applicable to different medical institutions. The experimental results show that the method in this paper not only improves the accuracy of the model detection, but also verifies its good generalization effect on different datasets.

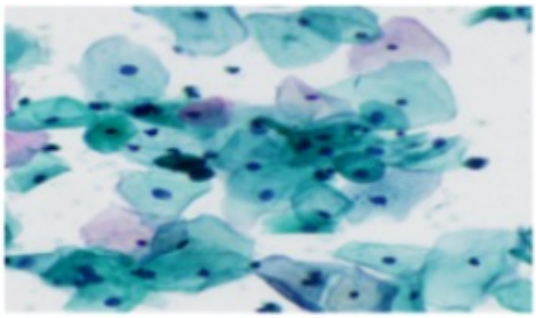
Keywords: Cervical Fluid-based Cell Detection, RetinaNet, Causal Autocoder, Deformable Convolution, Attention Mechanism

1. Introduction

Cervical cancer is the most common gynecologic malignancy and has the 3rd highest incidence rate among female malignancies worldwide Arbyn et al. (2011). In clinical practice, cervical cancer is best treated and requires lower medical costs if it is detected at an early stage Bray et al. (2018). Thinprep cytologic test (TCT) is the best early detection method for cervical cancer as determined by the WHO Organization et al. (2006). Reading cervical pathology images is an extremely labor-intensive, complex and tedious task. Each pathology image has thousands of cells and the pathologist needs to scan carefully and make a judgment. If pathologists are not skilled or overworked they are prone to an increased rate of misdiagnosis and underdiagnosis William et al. (2019). For instance, the estimated false-negative rates vary widely, ranging from 2% to 55% Branca and Longatto-Filho (2015).



(a) CDTBS Dataset



(b) A Certain Hospital Dataset

Figure 1: Examples of different cervical cell datasets. Figure 1(a)subfigure is a publicly available CDTBS dataset and Figure 1(b)subfigure is a clinically collected dataset from a hospital. They have problems such as image fading and insufficient cell staining that can seriously affect the performance of the model on different datasets.

In recent years, deep learning techniques have been heavily applied to medical images Krizhevsky et al. (2017); Simonyan and Zisserman (2014); Ronneberger et al. (2015); Girshick et al. (2014); Liu et al. (2022); Yi et al. (2020); Ghoneim et al. (2020); Wang et al. (2022) due to their excellent results in all computer vision tasks.Menglu Zhang et al Zhang and Shen (2021) proposed an automatic cervical cytology analysis detection framework with effective feature representation, employing elastic transformation and channel space attention modules to obtain more powerful feature extractors, suppress unnecessary features, and effectively improve the accuracy of cell detection. Yixiong Liang et al Liang et al. (2022) used the contextual relationships between cells and between cells and global images to enhance each Region of Interest (RoI) feature and improve the classification accuracy. Zhonghua Peng et al Peng et al. (2023) used YOLOv5 as a baseline model and used Transformer Block combined with a multi-headed self-attentive layer to enhance feature extraction to better extract cell features and obtain global information. Dongyao Jia et al Jia et al. (2022) proposed an improved SSD network by adding supplementary cell features to improve the overall accuracy of the model. Tingting Chen et al Chen et al. (2022) improved the detection performance by learning more effective feature representations for specific cell structures. These methods take full advantage of the fact that deep learning

can automatically extract complex features from data at multiple levels, and automatically learn and extract morphology, color, texture, and other features from cell images for cell detection tasks, achieving better detection results.

However, due to the difficult and time-consuming acquisition and labeling of medical image data, models are usually trained and evaluated on datasets collected from a single medical institution. Thus the models have poor generalization capabilities, resulting in poor application to other datasets. Deep learning models extract the high-order correlation properties or features for pattern analysis or synthesis Deng (2014) Sarker (2021). Yet, these features are unstable across cross-domain datasets, which can affect the generalization ability and detection effectiveness of the model. For example, background features have strong correlation with the goal cells, but the different equipment and staining methods used by different medical institutions lead to differences in image quality and feature stability across different datasets, as shown in Figure 1. Meanwhile, the problem of diverse shape and size variations of cervical cells themselves can also seriously affect the generalization performance of the model.

In order to improve the generalization of the model, this paper proposes a new method for cervical fluid-based cell detection based on causal invariance theory, which ensures good generalization of the model to new datasets by extracting potentially invariant causal feature representations from cervical fluid-based pathology image data. The main contribution points include:

(1) To address the interference problem of task-irrelevant features, a causal autoencoder is introduced to extract the causal feature representations of cells. Focuses on identifying key robust invariant features of different classes of cervical cells and reduces the impact of task-irrelevant feature representations in the data. The effect of the variability of feature distribution in different datasets is reduced, and the domain shift problem is effectively solved;

(2) To address the interference problem of different background categories of cervical fluid-based pathology images, deformable convolution and attention mechanisms are added. On the one hand, the deformable convolution enhances the learning ability of the model for different cellular irregular contour features. On the other hand, the attention mechanism enhances the feature extraction ability of the model for different foreground categories with variable shapes in the cervical fluid-based pathology images. By reducing the impact of possible strong correlation between background features and goal cells, the interference of foreground categories by problems such as fading and not bright enough staining of stained films is reduced. The generalization ability of the model is improved, which allows the model to be better applied to different medical institutions.

2. Methods

Figure 2 shows the overall architecture of the model, we use RetinaNet Lin et al. (2017) as the baseline model, and first replace the normal convolution with Deformable Convolution Block (DC-Block) in the backbone network ResNet-50, so that the model does not focus only on the features in a rectangular box during the feature extraction process. Then, we add the Spatial Attention-based interpretable module (SA), which makes the model focus

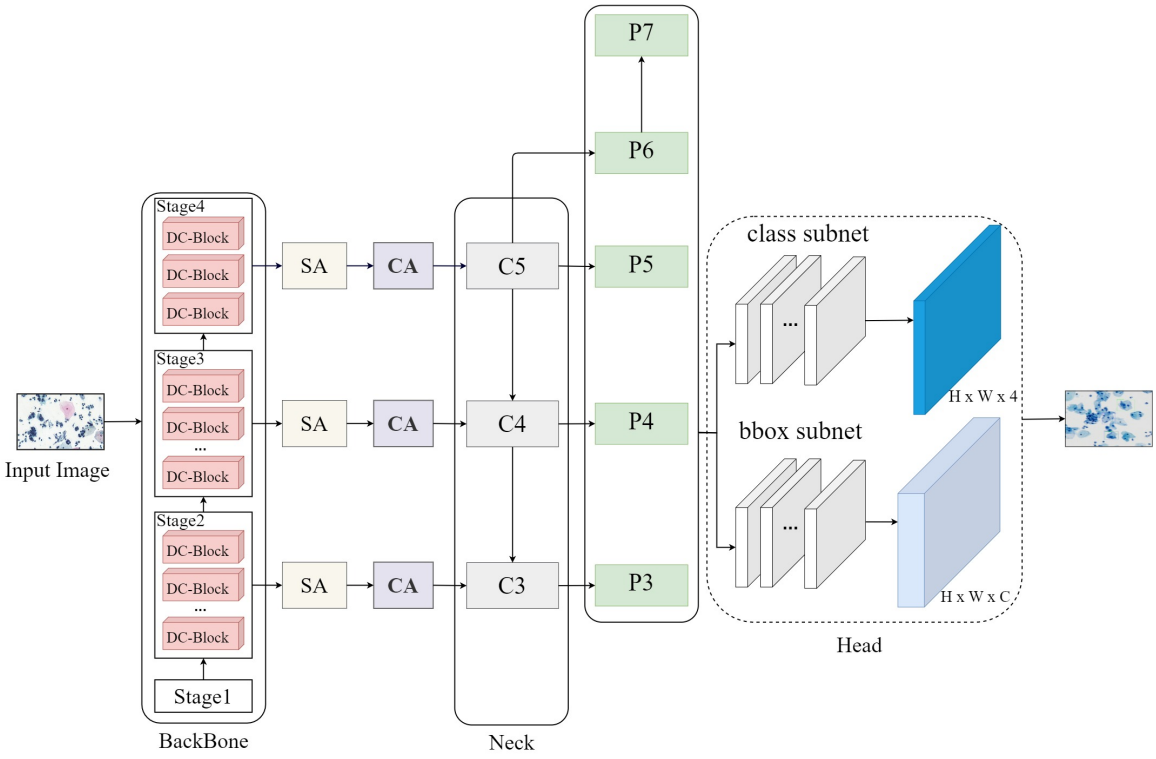


Figure 2: The general architecture of the model. Where DC-Block denotes the Deformable Convolutional block, SA denotes the Spatial Attention module, and CA denotes the Causal Autoencoder module. In order to reduce the computational resources occupied by feature maps, RetinaNet starts from C3 for subsequent computation, and the neck network FPN((Feature Pyramid Network) receives three feature maps C3, C4 and C5, and outputs five feature maps P3-P7 into the head network for classification and regression.

on the foreground class to be detected and reduce the attention to the background class in the feature extraction process. Causal Autoencoder (CA) is used to extract robust invariant causal features for identifying different classes of cells, thus enhancing the generalization ability of the model. The implementation of each module is described separately in the following.

2.1. Deformable Convolution Block

Usually, when the convolution operation is performed on the feature map, the result value of each pixel point after convolution is equal to the cumulative sum of the product of the pixel point values within the size of the convolution block and the corresponding convolution block position weights, centered on the current pixel point. The calculation is shown in Equation 1,

$$y\left(p_i\right)=\sum_{p_j \in \text{ kernal}} w\left(p_j\right) \cdot x\left(p_i+p_j\right) \tag{1}$$

where p_i and p_j enumerate the locations in $kernel$, specifically, p_i represents the current location to be calculated, the $kernel$ defines the receptive field size and dilation. For example, $kernal = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$ defines a 3×3 kernel with dilation 1. And $y(p_i)$ represents the feature value of p_i in the feature map after convolution, w represents the convolution kernel, and $w(p_j)$ represents the weight value corresponding to p_j in the convolution kernel.

This convolution is limited by the size of the convolution kernel, the extent of the perceptual field is basically equal to the size of the convolution kernel, and the features that can be captured by the feature extraction module are thus limited to consider only pixels of the rectangular convolution block around the convolution point. However, cervical cells are usually in shapes such as circles, ovals and irregular polygons, and there is no regularity in the size of the cells. In cervical cell detection, the complexity of variable size of different cervical cell shapes can seriously affect the performance of the model. In order to avoid the performance impact of incomplete target feature extraction, we add deformable convolution to the model to replace the normal convolution block to extract target features of different sizes and shapes, and then optimize the model.

The deformable convolution Dai et al. (2017) adds an additional offset clause Δp_j to the traditional standard convolution in the calculation of the sampling position, and the kernel is augmented with offsets $\{\Delta p_j \mid n = 1, ..., N\}$, where $N = |kernel|$, as shown in Equation 2.

$$y(p_i) = \sum_{p_j \in \text{ kernal}} w(p_j) \cdot x(p_i + p_j + \Delta p_j) \tag{2}$$

The offset map calculated by offset convolution has twice the number of channels as the size of the convolution kernel, and stores the offset of each pixel point on the feature map for each position in the convolution kernel in the horizontal and vertical coordinates, respectively. During the training process, the parameters of the offset term can also be continuously trained and make the convolution process capture the features of the cell itself more accurately, reduce the influence of surrounding noise features, and improve the model feature extraction ability.

2.2. Interpretable Modules Based On Spatial Attention

In addition to the effect of the shape of the cells themselves, the number of instances of the background category in the cervical cell images (e.g., the presence of a large number of partial adenocarcinomas with squamous intraepithelial neoplasia) is much larger than the instances of the foreground category. Moreover, a certain amount of background noise is generated during the production of pathology images, and these issues also have an impact on the performance of the model. To solve these problem, we introduce a spatial attention mechanism to improve the model's recognition of foreground categories as a way to improve the model generalization ability.

The spatial attention mechanism Woo et al. (2018) compresses the feature information of multiple channels in a feature map to finally obtain an attention map with dimension 1 and the same size as the input features, which responds to the importance of each pixel location in space. Specifically, the spatial attention mechanism is used to calculate the channel average and maximum value of each pixel of the input feature map F and superimpose

them to generate new features. The feature map with channel dimension 1 is generated by 7×7 convolution, and after *Sigmoid* activation, a spatial attention map with values ranging from 0 to 1 is obtained, indicating the importance of each pixel position in space, as shown in Equation 3,

$$M_s(F) = \sigma\left(f^{7\times7}\left(\left[AvgPool(F); MaxPool(F)\right]\right)\right) \tag{3}$$

where $f^{7\times7}$ represents a convolutional block calculation of 7×7, $AvgPool()$ and $MaxPool()$ represent average pooling and maximum pooling calculations on the input feature map, respectively, and then concatenate the two pooling results as inputs to the function $f^{7\times7}$, σ represents a *Sigmoid* operation.

The new feature map F' is obtained by multiplying the obtained spatial attention map $M_s(F)$ with the original input feature map F, as shown in Equation 4.

$$F' = M_s(F) \otimes F \tag{4}$$

2.3. Causal Autoencoder

Although the deformable convolution block and spatial attention mechanisms have enabled the model to focus on all features of the foreground cell classes as much as possible. However, not all features are useful when actually distinguishing between different classes of cells. For example, cells of the same class in different datasets may differ in features such as color and texture depending on the medical institution, and these features are not the basis for distinguishing between cell classes, but rather affect the ability of the model to generalize across different datasets. In contrast, pathologists typically identify abnormal cells based on nucleus size and shape, features that are stable and invariant across domains and thus require special attention from the model. To address this problem, we implemented a causal autoencoder to solve it.

High-dimensional data tends to reduce the performance of feature learning, and autoencoder can extract useful information from the high-dimensional data of cervical pathology image data by unsupervised learning, thus reducing the data dimensionality. However, the commonly used autoencoder extracted information will have redundant information with strong correlation with the goal cells, which severely reduces the generalization ability of the model. Therefore, this paper uses causal autoencoder to perform feature extraction and extract feature information with causal relationship with the goal cells. As shown in Figure 3, the causal autoencoder learns the low-dimensional representation by minimizing the reconstruction error between the input and output data, and divides the low-dimensional features into two categories by causal structure learning, i.e., causal features and other irrelevant features.

The convolutional autoencoder consists of two stages, encoding and decoding, where given the input data X, the encoder encodes it to learn the low-dimensional feature representation ξ, and then the decoder decodes it to obtain the estimated output ψ. The computational process is represented as follows,

$$\text{Encoding: } \xi^{(j)} = \sigma\left(\text{Conv}\left(\xi^{(j-1)}, \mathbf{W}_1^{(j)}\right) + \mathbf{b}_1^{(j)}\right), j = 1, 2, \cdots, l. \tag{5}$$

$$\text{Decoding: } \psi^{(j)} = \sigma\left(\text{Deconv}\left(\psi^{(j-1)}, \mathbf{W}_2^{(j)}\right) + \mathbf{b}_2^{(j)}\right), j = 1, 2, \cdots, l. \tag{6}$$

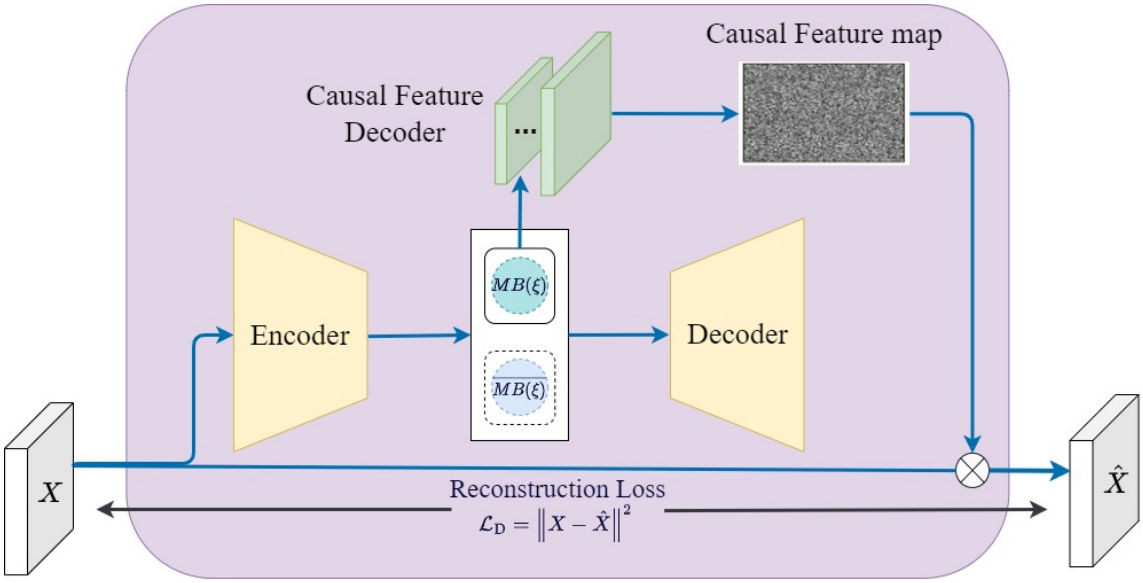


Figure 3: Architecture diagram of causal autoencoder. It consists of two parts: a convolutional autoencoder maps the input data to a low dimensional feature space to reduce the impact of noise in the data, and a causal feature decoder remaps the learned causal features to a high dimensional feature space.

where l is the number of convolutional layers of the convolutional autoencoder, σ is the nonlinear activation function, $\mathbf{W}$ is the weight matrix, and $\mathbf{b}$ is the bias vector.

Since the low-dimensional features learned by the convolutional autoencoder may contain some redundant information that is not relevant to the goal cells. For example, the background of cervical cell images and other irrelevant cells, which can reduce the robustness of the model. Therefore, we expect that the model can focus on learning feature representations that are causally related to the goal cells, since these causal features are stable and invariant across domains.

Markov Blanket (MB) feature selection methods can quickly remove redundant features and obtain the optimal feature subset Guyon and Elisseeff (2003); Ling et al. (2019); Yu et al. (2020). Under the Markov Blanket condition of goal features, the valid information of the goal is contained in its Markov Blanket, and all non-Markov Blanket ($\overline{MB}$) features are regarded as redundant features. We consider the MB features of the goal cell as its causal features and the $\overline{MB}$ features as irrelevant features. To obtain the MB features of the goal variable, the directed acyclic graph G on the low-dimensional feature ξ is first learned. Once G is obtained, the low-dimensional features obtained by encoding can be distinguished into causal features and task irrelevant features.

Inspired by Zheng et al. (2018), we use the adjacency matrix $A = [a_1 \mid ... \mid a_k] \in \mathbb{R}^{k \times k}$ to encode the directed acyclic graph of ξ, where each column a_i denotes the coefficients of the linear structural equation model $Z_i = a_i^T \xi + N_i$, N_i denotes the additive noise, and $Z_i \in Z$ is a parent (a direct cause) of Z_j and Z_j is a child (a direct effect) of Z_i if there is a directed edge from variable Z_i to variable Z_j, and Z is the set of k-dimensional high-level

feature representation variables. Under the condition that A is acyclic Yang et al. (2021), a new matrix $\hat{A}$ is obtained by thresholding the edge weights of A,

$$\hat{A}[i][j] = \begin{cases} 0, & if|A[i][j]| < \sigma \\ A[i][j], & \text{otherwise.} \end{cases} \tag{7}$$

if $|A\,[i]\,[j]| \geq \sigma$ holds, it means that there exists a directed edge from variable Z_i to variable Z_j, i.e., Z_i is a direct cause of Z_j.

Because in a directed acyclic graph, the Markov Blanket of a node is the parent node, child node and parent of child node of that node. So once $\hat{A}$ is obtained, we can find these nodes with causal relationship with the goal variable by matrix $\hat{A}$ to identify the MB feature representation of the goal cell, and we can divide the low-dimensional features into two groups, which are causal features $MB\,(\xi)$ and other irrelevant features $\overline{MB\,(\xi)}$. Then we use the causal feature decoder to decode the learned low-dimensional causal features, and the obtained causal feature graph is computed with the input feature graph to obtain the new feature graph.

We perform unsupervised pre-training on it using the CDTBS training dataset. The overall loss function is shown in Equation 8,

$$\mathcal{L}_C = \mathcal{L}_D + \alpha\mathcal{L}_R \tag{8}$$

where $\mathcal{L}_D$ denotes the reconstruction loss, $\mathcal{L}_R$ is the model parameter regularization term, and α is used as the weight value to adjust the loss weight.

2.4. Network Optimization

The features output by the causal autoencoder are input to the FPN, and then the multi-scale image features are output and fed into two prediction modules, the classification subnet and the regression subnet, respectively. The probability of which class the cell belongs to is predicted in the classification subnet, and the location of the cell is predicted in the regression subnet, and finally the two predictions are combined to obtain the cell detection result. In the classification subnet, Focal Loss is used to calculate the classification loss $\mathcal{L}_{cls}$ as shown in Equation 9,

$$\mathcal{L}_{cls} = -\alpha_t\,(1-p_t)^{\gamma}\log\,(p_t) \tag{9}$$

In the regression subnet, $\mathcal{L}_{bbox}$ represents the coordinate offset regression loss of the anchor frame generated according to the anchor point, and Smooth L1 loss is used in this paper instead of L1 loss for the calculation, as shown in Equation 10,

$$\mathcal{L}_{\text{bbox}} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{10}$$

where x represents the difference between the predicted value and the true label. When the prediction difference is less than 0.5, L2 loss is used for calculation, conversely, a certain degree of offset is added on the basis of L1 loss calculation. Thus Smooth L1 loss inherits the advantages of both, solves the gradient explosion problem during training, and improves the stability of training. Thus the loss function of the final goal detection model is calculated as follows,

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{bbox} \tag{11}$$

3. Experiments

3.1. Datasets

To evaluate the model performance, two datasets were selected: the CDTBS Liang et al. (2018) and a clinical cervical fluid-based pathology image dataset from a hospital:

CDTBS Dataset. It is a public dataset. The dataset contains 7410 cervical fluid-based pathology images cropped from Whole-Slide Images (WSI) obtained from digital scanners, containing a total of 11 cell categories, which are ASC-US (ascus), ASC-H (asch), low-grade squamous intraepithelial lesion (lsil), high-grade squamous intraepithelial lesion (hsil), squamous-cell carcinoma (scc), atypical glandular cells (agc), trichomonas (trich), candida, flora, herps and actinomyces (actin).

A Certain Hospital Dataset. It was acquired from clinical cervical cancer screening in a hospital and desensitized. The dataset contains 11,739 cervical fluid-based pathology images cropped from WSI obtained from a digital scanner, of which 3,171 were labeled by pathologists, and the test dataset contains 518 labeled images. A total of 8 categories of cells are included, which are asch, ascus, hsil, scc, agc, lsil, candida and flora.

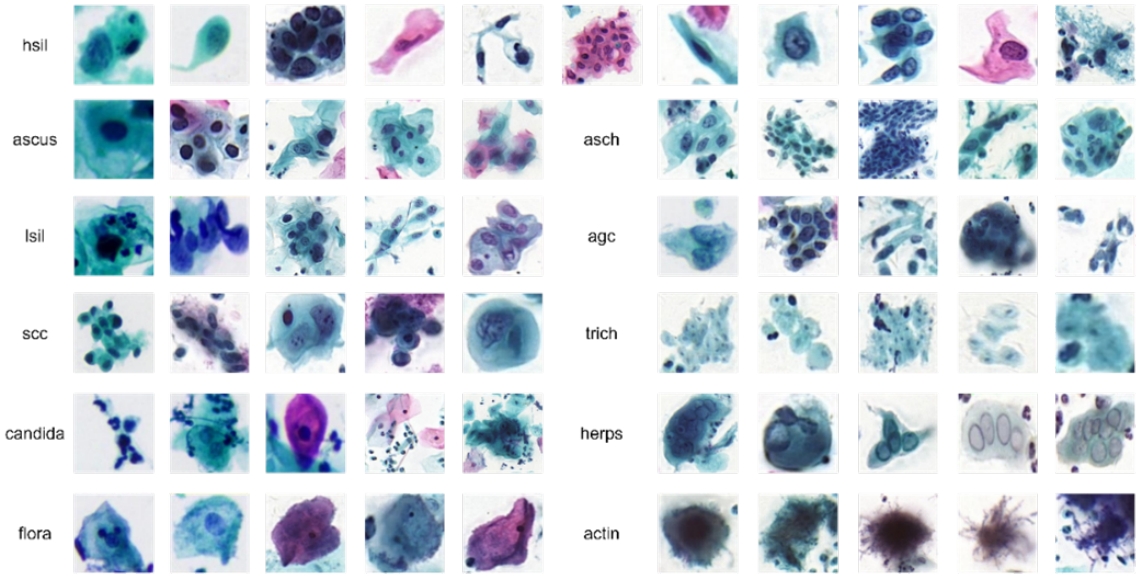


Figure 4: Example of cervical cells.

Data Augmentation. For an original image of the training dataset, three different clockwise rotations of 90 degrees, 180 degrees and 270 degrees are performed first. Then they are flipped horizontally to obtain a new sample. After such two layers of data augmentation, compared with the original training dataset, the cell size and angle transformations are more abundant, and the overall sample size is increased while the sparse categories are supplemented in large quantities, enabling the model to fully learn the different category features.

3.2. Experimental Setup and Evaluation Standard

Implementation Details. The causal autoencoder is pre-trained using the Adam Kingma and Ba (2014) optimizer with the batchsize set to 2 and the learning rate set to 0.001. In Equation 8, the parameter α is set to 0.01. The backbone network of the overall model is ResNet-50 obtained by pre-training on ImageNet Deng et al. (2009), the neck feature extraction network is FPN, and the feature dimension of the output to the prediction module is 256 dimensions. The input cervical cell images are uniformly 1024 $\times$ 1024 pixel size, using an SGD Zhou et al. (2020) optimizer with a learning rate of 0.0025, momentum set to 0.9, weight decay to 0.0001, and batchsize set to 8. The overall model is implemented through the PyTorch deep learning framework, and the hardware is configured with a 1TB hard disk and 128G of memory. The CPU configuration is 8-core 2.40GHz*4, and the GPU configuration is NVIDIA GeForce 2080 Ti.

Assessment. The assessment metric we used was mAP (mean Average Precisions), which is the average AP value for each category. Usually, AP50 is considered to have indicative results for good localization and classification scores. Therefore, we use an IoU threshold of 0.5 to calculate AP and then combine the average as the final evaluation result.

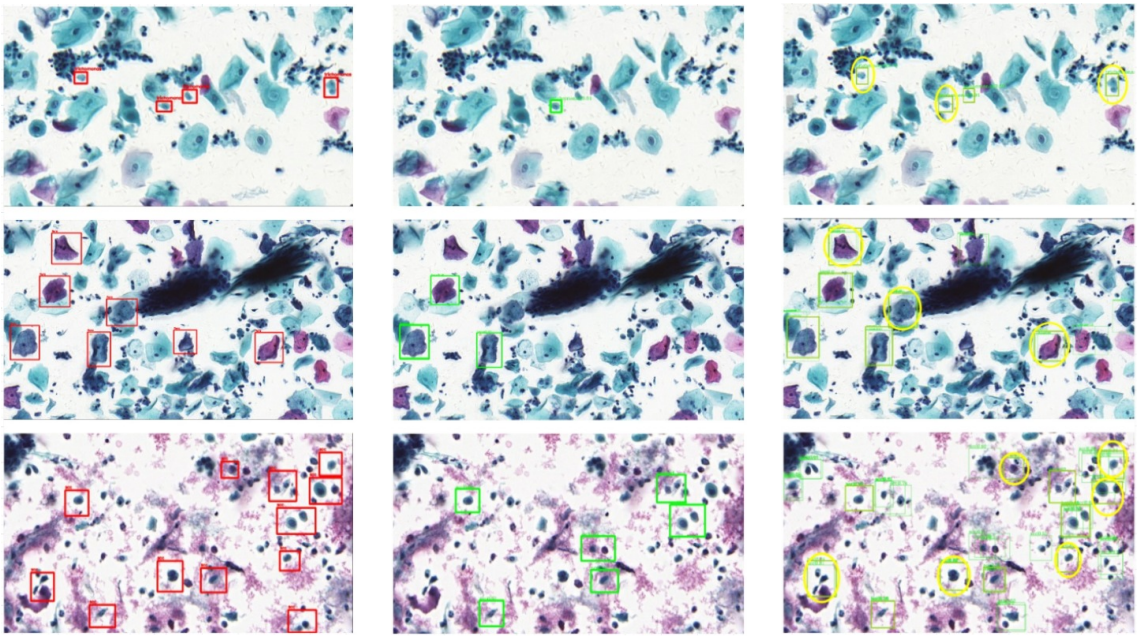
3.3. Experimental Results

Analysis Of Single-domain Experimental Results. The single-domain experiments were trained on the training dataset of CDTBS and tested on the test dataset of CDTBS. The results of our model were compared with other classical one-stage models, and the results are shown in Table 1. As can be seen from the table, our proposed model has a +2.3% improvement in mAP compared to the baseline RetinaNet model, and a +0.9% and +1.1% improvement compared to SSD Liu et al. (2016) and FCOS Tian et al. (2019), respectively. In addition, our method also has better results compared to other cervical cell detection models. Compared with the detection accuracy of the model Comparison Detector Liang et al. (2018) using the same dataset for experiments, our model has a significant improvement in detection accuracy, which is +5.6% higher.

Table 1: Experimental results of different models on the CDTBS dataset

Method	BatchSize	mAP
SSD	8	50.6
FCOS	8	50.4
RetinaNet	8	49.2
Comparison Detector	4	45.9
Ours	8	51.5

The detection of cervical cancer cells and precancerous cells is crucial as missed diagnoses in cervical fluid-based cell testing can affect the timely treatment of patients, and morbidity and mortality rates can increase as a result. A visual analysis of the RetinaNet model and our model test results is presented in Figure 5. From the figure, it can be found that compared with RetinaNet, our model can identify abnormal cells more accurately and improve the detection rate, which is important for reducing morbidity and mortality.



(a) Ground Truth (b) Prediction of RetinaNet (c) Prediction of Ours

Figure 5: Visualization of prediction results. Figure 5(*a*)subfigure shows the ground truth annotation results, Figure 5(*b*)subfigure shows the cell detection results of RetinaNet drawn in green boxes, and Figure 5(*c*)subfigure shows the detection results of our model, with the yellow oval circles pointing out the abnormal cells detected by our model more than RetinaNet.

Ablation Experiments. The ablation experiments are used to verify the effectiveness of the modules of our model such as causal autoencoder, deformable convolutional block and spatial attention mechanism, and the experimental results are shown in Table 2.

As can be seen in Table 2, the first row is not good using only the baseline model. The second row is a +15.4% improvement in mAP after data augmentation, which indicates that the cell sizes and angles become richer after rotation and flip, enhancing the applicability of the model to the unknown transformation of the goal. The third row is a +1.3% improvement compared to the baseline model after adding the causal autoencoder, indicating that the causal features effectively reduce the impact of task-irrelevant features and help the model perform classification detection more than the correlation features. The fourth row shows a +0.2% improvement in mAP after replacing the normal convolution block with the deformable convolution block, indicating that the deformable convolution is helpful for the model to detect various types of cells with different morphologies. The fifth row is the introduction of the spatial attention mechanism, which gives the model a +1.9% boost, a good proof that it can better capture the foreground category and reduce the interference of the background category. The last row shows the result of our model, which has a significant boost of +2.3% over RetinaNet.

Table 2: Ablation experimental results of the model on the CDTBS dataset (DA: Data Augmentation; CA: Causal Autoencoder; DC: Deformable Convolution; SA: Spatial Attention)

DA	CA	DC	SA	mAP
				33.8
✓				49.2
✓	✓			50.5
✓		✓		49.4
✓			✓	51.1
✓	✓	✓	✓	51.5

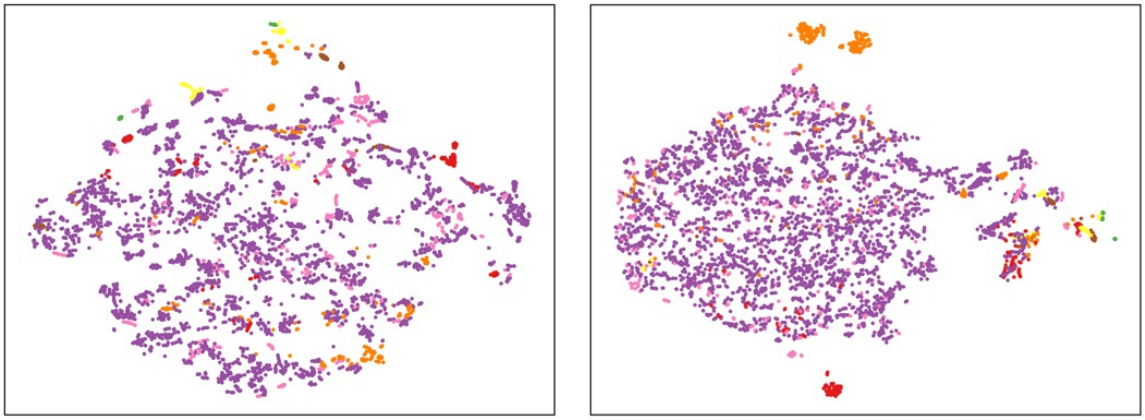
Cross-domain Generalization Capability Assessment. In the field of machine learning, the generalization ability of a model usually refers to its performance on new data. Models with good generalization ability are able to perform well on unseen data, while models with poor generalization ability may perform poorly on new data. To verify that our method can effectively improve the generalization ability of the model, we selected two different datasets for cross-domain experiments: the CDTBS dataset was used as the source domain to train the model and tested on a clinical cervical cancer screening pathology image dataset from a hospital. In this experiment, we unified the annotations of the CDTBS dataset into the same eight categories as a hospital dataset, and removed the redundant images that did not contain these categories from the training dataset. The experimental results are shown in Table 3.

Table 3: Results of cross-domain experiments

Method	mAP
RetinaNet	38.8
RetinaNet+CA	41.3
RetinaNet+DC	40.4
RetinaNet+SA	40.2
Ours	42.7

The experimental results show that the performance of the model on the cross-domain experiments is improved by +2.5%, +1.6% and +1.4% after adding the causal autoencoder, the deformable convolution and the spatial attention mechanism, respectively. Among them, the causal autoencoder improves the generalization ability of the model the most, which indicates that the causal autoencoder can better capture the actual causal relationships in the data, filter out the task-irrelevant features, reduce the influence of the variability of feature distribution among different datasets, and thus improve the robustness of the model. When all three methods are applied to the baseline model simultaneously, the mAP value improves by 3.9%, which effectively verifies that our method not only improves the prediction performance of the model, but also better improves the generalization ability of the model.

Further, we used t-SNE Van der Maaten and Hinton (2008) to visually validate the algorithm effectiveness. t-SNE is a dimensionality reduction technique that preserves local



(*a*) RetinaNet (*b*) Ours

Figure 6: Visualization of t-SNE. Each color in the figure represents a cell category, for a total of 8 categories.

relationships in high-dimensional data and maps high-dimensional data to 2D or 3D space for visualization purposes. Specifically, it calculates the distance between each sample point and its adjacent sample points, converts these distances into probability distributions to represent their similarity with other data points, and then uses Gaussian kernel functions to map these probability distributions into 2D or 3D space. RetinaNet and our method performed the cell detection on the CDTBS dataset separately, and the visualization of the t-SNE data distribution of the detection results is shown in Figure 6.

It can be seen that the data distribution of the same category in Figure 6(*a*)subfigure is relatively discrete and the features between different categories stick to each other severely, while in Figure 6(*b*)subfigure relative to Figure 6(*a*)subfigure the same category becomes close to each other and the spacing with different categories becomes larger. It can be seen that our method facilitates the alignment of category features, which makes the model easier to classify and thus improves the generalization ability of the model.

4. Conclusion and Outlook

In this paper, a cervical fluid-based cell detection model with strong generalization capability is proposed based on RetinaNet. A causal autoencoder is used to learn the causal features of goal cells in pathological images, while a spatial attention mechanism and deformable convolution are introduced to improve the grasping ability of foreground categories. The comparison of experimental results shows that our method improves the accuracy of the baseline model and outperforms other classical models while improving the accuracy of the baseline model by 3.9% in cross-dataset experiments, which greatly improves the cross-domain detection capability of the model. In the future, we will try more algorithms to learn the causal structure in the cervical fluid-based cell images and optimize the network structure and training method of the causal autoencoder to make it better applied to cervical abnormal cell detection. And we have also noticed that advanced data augmentation techniques in pathology may improve model generalization performance, and we will focus

on it as a research focus in the future. In addition, we will quantitatively compare the efficiency impact of each module through experimental analysis in the future, seeking a better balance between efficiency and accuracy.

References

Marc Arbyn, Xavier Castellsagué, Silvia de Sanjosé, L Bruni, M Saraiya, F Bray, and J Ferlay. Worldwide burden of cervical cancer in 2008. *Annals of oncology*, 22(12): 2675–2686, 2011.

Margherita Branca and Adhemar Longatto-Filho. Recommendations on quality control and quality assurance in cervical cytology. *Acta cytologica*, 59(5):361–369, 2015.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, Ahmedin Jemal, et al. Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca Cancer J Clin*, 68(6):394–424, 2018.

Tingting Chen, Wenhao Zheng, Haochao Ying, Xiangyu Tan, Kexin Li, Xiaoping Li, Danny Z Chen, and Jian Wu. A task decomposing and cell comparing method for cervical lesion cell detection. *IEEE Transactions on Medical Imaging*, 41(9):2432–2442, 2022.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA transactions on Signal and Information Processing*, 3:e2, 2014.

Ahmed Ghoneim, Ghulam Muhammad, and M Shamim Hossain. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, 102:643–649, 2020.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Dongyao Jia, Zihao He, Chuanwang Zhang, Wanting Yin, Nengkai Wu, and Ziqi Li. Detection of cervical cancer cells in complex situation based on improved yolov3 network. *Multimedia Tools and Applications*, 81(6):8939–8961, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Yixiong Liang, Zhihong Tang, Meng Yan, Jialin Chen, Qing Liu, and Yao Xiang. Comparison-based convolutional neural networks for cervical cell/clumps detection in the limited data scenario. *arXiv preprint arXiv:1810.05952*, 2018.

Yixiong Liang, Shuo Feng, Qing Liu, Hulin Kuang, Jianfeng Liu, Liyan Liao, Yun Du, and Jianxin Wang. Exploring contextual relationships for cervical abnormal cell detection. *arXiv preprint arXiv:2207.04693*, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Zhaolong Ling, Kui Yu, Hao Wang, Lin Liu, Wei Ding, and Xindong Wu. Bamb: A balanced markov blanket discovery approach to feature selection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–25, 2019.

Minmin Liu, Xuechen Li, Xiangbo Gao, Junliang Chen, Linlin Shen, and Huisi Wu. Sample hardness based gradient loss for long-tailed cervical cell detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 109–119. Springer, 2022.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

World Health Organization, World Health Organization. Reproductive Health, World Health Organization. Chronic Diseases, and Health Promotion. *Comprehensive cervical cancer control: a guide to essential practice*. World Health Organization, 2006.

Zhonghua Peng, Rong Hu, Fuen Wang, Haoyi Fan, Yee Wei Eng, Zuoyong Li, and Liwei Zhou. Deep adaptively feature extracting network for cervical squamous lesion cell detection. In *International Conference on Machine Learning for Cyber Security*, pages 238–253. Springer, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wei Wang, Yun Tian, Yang Xu, Xiao-Xuan Zhang, Yan-Song Li, Shi-Feng Zhao, and Yan-Hua Bai. 3cde-net: a cervical cancer cell detection network based on an improved backbone network and multiscale feature fusion. *BMC Medical Imaging*, 22(1):130, 2022.

Wasswa William, Andrew Ware, Annabella Habinka Basaza-Ejiri, and Johnes Obungoloch. A pap-smear analysis tool (pat) for detection of cervical cancer from pap-smear images. *Biomedical engineering online*, 18:1–22, 2019.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Lin Yi, Yajie Lei, Zhichen Fan, Yingting Zhou, Dan Chen, and Ran Liu. Automatic detection of cervical cells using dense-cascade r-cnn. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part II 3*, pages 602–613. Springer, 2020.

Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5):1–36, 2020.

Menglu Zhang and Linlin Shen. Cervical cell detection benchmark with effective feature representation. In *Cognitive Systems and Signal Processing: 5th International Conference, ICCSIP 2020, Zhuhai, China, December 25–27, 2020, Revised Selected Papers 5*, pages 402–413. Springer, 2021.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. *Advances in Neural Information Processing Systems*, 33:810–821, 2020.