Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT

Chen Zhao 1 · Renjun Shuai 1 · Li Ma 2 · Wenjia Liu 3 · Menglin Wu 1

Abstract
Cervical cell classification has important clinical significance in cervical cancer screening at early stages. However, there are fewer public cervical cancer smear cell datasets, the weights of each classes’ samples are unbalanced, the image quality is uneven, and the classification research results based on CNN tend to overfit. To solve the above problems, we propose a cervical cell image generation model based on taming transformers (CCG-taming transformers) to provide high-quality cervical cancer datasets with sufficient samples and balanced weights, we improve the encoder structure by introducing SE-block and MultiRes-block to improve the ability to extract information from cervical cancer cells images; we introduce Layer Normalization to standardize the data, which is convenient for the subsequent non-linear processing of the data by the ReLU activation function in feed forward; we also introduce SMOTE-Tomek Links to balance the source data set and the number of samples and weights of the images we use Tokens-to-Token Vision Transformers (T2T-ViT) combing transfer learning to classify the cervical cancer smear cell image dataset to improve the classification performance. Classification experiments using the model proposed in this paper are performed on three public cervical cancer datasets, the classification accuracy in the liquid-based cytology Pap smear dataset (4-class), SIPAKMed (5-class), and Herlev (7-class) are 98.79%, 99.58%, and 99.88%, respectively. The quality of the images we generated on these three data sets is very close to the source data set, the final averaged inception score (IS), Fréchet inception distance (FID), Recall and Precision are 3.75, 0.71, 0.32 and 0.65 respectively. Our method improves the accuracy of cervical cancer smear cell classification, provides more cervical cell sample images for cervical cancer-related research, and assists gynecologists to judge

* Renjun Shuai
srjwhy@njtech.edu.cn

1 College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China
2 Nanjing Health Information Center, Nanjing 210003, China
3 Changzhou No. 2 People’s Hospital affiliated with Nanjing Medical University, Changzhou 213003, China
and diagnose different types of cervical cancer cells and analyze cervical cancer cells at different stages, which are difficult to distinguish. This paper applies the transformer to the generation and recognition of cervical cancer cell images for the first time.

Keywords  Cervical cancer · Cervical cytology · Cervical cell classification · Convolutional neural networks · Generative adversarial networks · Vision transformer

1 Introduction

Cervical cancer is the fourth most common cause of death from cancer in females [48]. It is estimated that there will be 604,127 cases and 341,831 deaths worldwide in 2020, and it is the second most common cancer in women worldwide [28, 52]. At early stages of cervical cancer, the cure rate is nearly 100% [42], so the prevention, early detection and classification of cervical cancer are essential [64].

At present, cervical cancer screening methods mainly include human papillomavirus detection, cervical smear and acetic acid testing under colposcopy [72]. After the introduction of a Papanicolaou (Pap) smear [40], the standard screening test for cervical cancer and premalignant lesions is cervical cytology. As the most common screening test, cervical cytology has been extensively used and effectively reduces incidence and mortality. At present, manual screening of abnormal cells from a cervical cytology slide is still common practice. However, it is usually tedious, inefficient and expensive. Consequently, automatic screening methods have attracted increasing attention [3, 5]. Additionally, some research on cervical cell analysis shows that each independent cervical cell has intrinsic similarity. For example, superficial and intermediate cells generally have relatively small nuclei and have clear cytoplasmic and nuclear margins, while dyskeratotic and metaplastic cells have overlapping cytoplasmic and nuclear margins. In addition, koilocytotic cells have the presence of a perinuclear cavity, while other cells have a relatively thick cytoplasm [18, 34]. These observations indicate that there exists a potential relationship between cervical cell images. Therefore, accurate cervical cell classification is crucial to the automatic screening method. The analysis of Pap smear images requires low error tolerance and skilled pathologists, and the screening process is expensive and time-consuming. Therefore, an automated classification process can assist gynecologists in diagnosis and provide more objective test explanations.

Recently, deep learning has brought considerable improvements in accuracy in many applications [46]. Due to its high accuracy in many fields, deep learning has become the most advanced machine learning technology. Deep learning and CNNs have been successfully used in breast cancer detection [7], skin cancer recognition [73], and COVID-19 recognition and analysis [26]. Among them, there are many studies based on convolutional neural networks, and convolutional neural networks (CNNs) have been the standard for 3D medical image classification and segmentation. The convolutional operations used in these networks, however, inevitably have limitations in modeling the long-range dependency due to their inductive bias of locality and weight sharing [69]. To solve these problems, the transformer was created. Beginning at the end of 2020, transformer-based research has gradually increased. At present, some transformer-based research has surpassed CNN-based research in the fields of image classification, image detection, and image segmentation [33]. Dosovitskiy et al. [11] proposed a vision transformer (ViT) for the first time to be applied to image classification. They applied a method that does not focus on pixels but focuses on small areas of the image. They believe
that dependence on CNN is not necessary and that applying direct pure converters based on image patch sequences can perform image classification tasks well.

Carion et al. [4] proposed a new method (DEtection TRansformer or DETR) that views object detection as a direct set prediction problem. Their approach streamlines the detection pipeline, effectively removing the need for many hand-designed components such as a nonmaximum suppression procedure or anchor generation that explicitly encode the prior knowledge about the task. Xie et al. [69] proposed a novel framework that efficiently bridges a convolutional neural network and a transformer (CoTr) for accurate 3D medical image segmentation. Compared to CNN, the self-attention in the transformer can produce a more interpretable model, from which the attention distribution can be checked, and each attention head can learn to perform different tasks. The number of operations required to calculate the association between two positions does not increase with distance.

Research on the classification of cervical cancer cells is mostly carried out on two public datasets. Herlev [27] consists of 917 images of Pap smear cells classified carefully by cytotechnicians and doctors. Each cell is described by 20 numerical features, and the cells fall into 7 classes. SIPaKMeD [43] consists of 4049 annotated cell images. The cells are classified by expert cytopathologists into five different classes. In addition to the public dataset, there is nonpublic dataset research [68]. Currently, the number of samples in the two public datasets is limited, and the classification accuracy of various studies is above 90%, which tends to overfit, and the cell types in various datasets are inconsistent. The weights are not balanced, and the clarity and quality of the samples are uneven. Regarding the nonpublic dataset, cervical cancer researchers cannot obtain the data and can only conduct research on the limited public dataset, and the research progress is limited.

To solve the above problems, this paper proposes a cervical cell image generation model based on taming transformers (CCG-taming transformers) and a classification model based on taming transformers and Tokens-to-Token Vision Transformers (T2T-ViT) for the first time. The method proposed in this paper expands its dataset sample size, balances the number and weight of each type of cervical cell, and generates high-quality sample images. A new dataset (liquid-based cytology Pap smear dataset [24]) is introduced to provide more objective materials for the cervical cancer cell generation model in this paper, and T2T-ViT is used to further improve classification accuracy, provide more objective information for gynecologists, improve the efficiency of clinical-pathological diagnosis, detect the patient’s condition in time, and improve the survival rate of cervical cancer patients. Figure 1 is an overview of the method framework of the model in this paper.

The main contributions and novelty of this model are as follows:

- We propose a cervical cancer cell sample image generation model based on taming transformers (CCG-taming transformers), to our best knowledge, this is the first model combining CNNs and Transformer and applied to cervical cancer research.
- In CCG-taming transformers, we adjust the encoder structure of VQGAN in the taming transformers: 1) we introduce new convolutional structures MultiRes-block to better realize the extraction and analysis of the key information of the cervical cancer cell image by the encoder; 2) we introduce SE-block to alleviate the vanishing gradient problem in the neural network; 3) we introduce Layer normalization to enhance the discrimination ability of feature representations. This model mainly solves the problem that there are few public data sets in the current cervical cancer research, the samples of each class are very different, and the weights and quantities of various samples in the data sets are not
balanced. The cervical cancer cell images generated by this model can provide certain reference value for cervical cancer research.

- We introduce SMOTE-Tomek Links to balance the source data set and the number of samples and weights of the images we generate.
- We introduce T2T-ViT combing transfer learning to classify the cervical cancer cell image dataset, which can solve the problem that the classification model based on CNNs may lose the details of the feature map.

2 Related work

Artificial intelligence and deep learning play an important role in cell classification, medical image classification, generation and analysis [17, 59, 64]. As new technologies develop, they become cost-effective and less time-consuming. They are now more popular than traditional methods (such as Pap smears, colposcopy, and cervicography) [8]. These technologies have nothing to do with human experience. Although they cannot replace gynecologists for pathological evaluation, they can provide assistance for clinical diagnosis to a large extent, improve the diagnostic efficiency of gynecologists, and reduce the subjective components of diagnosis.

Several works classify and detect cervical cancer in the literature. Convolutional neural networks (CNNs) [35] have been proposed to automatically learn multilevel features through hierarchical deep architecture. Wieslander et al. [63] used ResNet for binary classification of benign and malignant cervical cells. Plissiti et al. [43] proposed the annotated cervical cell image dataset SIPaKMeD and applied CNN to classify five types of cervical cells. Gautam et al. [14] proposed a patch-based approach using CNN combined with transfer learning for the segmentation of nuclei in single-cell images. Ghoneim et al. [15] proposed a cervical cancer

Fig. 1 Flowchart of the generation and classification of cervical cancer cell images in this paper
cell detection and classification system based on convolutional neural networks (CNNs) and achieved 91.2% accuracy in the classification problem (7-class). William et al. [65] proposed contrast local adaptive histogram equalization for image enhancement. Cell segmentation was achieved through a trainable Weka segmentation classifier, and a sequential elimination approach was used for debris rejection. Wang et al. [58] proposed multiscale representation for scene classification, which is realized by a global–local two-stream architecture. They also [60] explored the attention mechanism and propose a novel endto-end attention recurrent convolutional network (ARCNet) for scene classification, their research has made outstanding contributions to the classification field. Although the neural network framework based on CNN has achieved good accuracy in the classification of cervical cancer smear cell images, the training requires more computing resources, and the number of network layers needs to reach a certain depth to capture the deep level of the sample image information. Compared with the transformer, in the CNN-based model, the number of operations required to calculate the association between two positions through convolution increases with distance. The number of operations required to calculate the association between two positions based on the self-attention in the transformer is independent of the distance.

Generative adversarial networks (GANs) [1] have been successfully used to synthesize human faces, landscapes and even medical images; they are mainly used to expand the size of datasets and balance the weight and number of samples in each category. Pollastri et al. [44] presented a novel strategy that employs DC-GAN to augment data in the skin lesion segmentation task. The proposed framework generates both skin lesion images and their segmentation masks, making the data augmentation process extremely straightforward. Karras et al. [31] proposed an alternative generator architecture for generative adversarial networks, borrowing from style transfer literature. The new architecture leads to an automatically learned, unsupervised separation of high-level attributes and stochastic variation in the generated images, and it enables intuitive, scale-specific control of the synthesis. Thuy et al. [54] proposed combining deep learning, transfer learning and generative adversarial networks to improve classification performance. Fine-tuning on the VGG16 and VGG19 networks was used to extract the well-discriminated cancer features from histopathological images before feeding them into the neural network for classification. Han et al. [20] proposed a 3D multiconditional GAN (MCGAN) to generate realistic/diverse nodules placed naturally on lung computed tomography images to boost sensitivity in 3D object detection. However, they ignore the potential relationships among cervical cell images during feature learning, and thus, may influence the representation ability of CNN features. Besides, GANs usually require large training datasets, which are often scarce in the medical field, and GANs are only applied to medical image synthesis at a relatively low resolution. In this paper, we use CCG-taming transformers and SMOTE-Tomek Links to balance the weight and number of samples between different categories, through this method, we can generate more samples of cervical cancer cells in different categories, providing more objective reference value for the research of cervical cancer.

Amazing results from transformer models on natural language processing have encouraged the vision community to study their research on computer vision problems [19]. Vaswani et al. [57] proposed the transformer network architecture for the first time, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Valanarasu et al. [56] proposed a gated axial attention model that extends the existing architectures by introducing an additional control mechanism in the self-attention module to train the model effectively on medical images, they also proposed a local-global training strategy (LoGo) to train the model
effectively on medical images with significant performance. Zhu et al. [74] proposed deformable DETR, whose attention modules only attend to a small set of key sampling points around a reference. Deformable DETR can achieve better performance than DETR (especially on small objects) with 10 fewer training epochs. He et al. [22] introduced DropBlock as a regularization technique for HSI accurate classification to mitigate the overfitting problem in CNN-based HSI classification.

Chen et al. [6] studied the low-level computer vision task (e.g., denoising, superresolution and deraining) and developed a new pretrained model, namely, an image processing transformer (IPT). Wang et al. [62] investigated a simple backbone network useful for many dense prediction tasks without convolutions, and they proposed the pyramid vision transformer (PVT), which overcomes the difficulties in applying the transformer to various dense prediction tasks. The transformer-based model has achieved high accuracy in image classification, target detection and other fields. Park et al. [41] proposed a novel vision transformer by using the low-level CXR feature corpus that is obtained to extract the abnormal CXR features. Wang et al. [61] combined transfer learning with the transformer model to predict the small-dataset Heck reaction. Besides, the simple tokenization of input images fails to model the important local structure (e.g., edges, lines) among neighboring pixels, leading to its low training sample efficiency, and the redundant attention backbone design of ViT leads to limited feature richness in fixed computational budgets and limited training samples. To solve these problems, we introduced T2T-ViT, which introduces 1) a layer-wise Tokens-to-Token (T2T) transformation to progressively structurize the image to tokens by recursively aggregating neighboring Tokens into one Token (Tokens-to Token), such that local structure presented by surrounding tokens can be modeled and tokens length can be reduced; 2) an efficient backbone with a deep-narrow structure for vision transformers motivated by CNN architecture design after extensive study.

3 Method

In this paper, we propose a cervical cell image generation model based on taming transformers (CCG-taming transformers) and a classification model based on Tokens-to-Token Vision Transformers (T2T-ViT) with transfer learning. The overall framework of the proposed model is illustrated in Fig. 2.

First, we used ImageNet to train T2T-ViT, preprocessed the cervical cancer Pap cell smear dataset, used CCG-taming transformers to expand the dataset, and then balanced the weights and numbers of different classes of images. After that, the newly obtained dataset was passed to the T2T-ViT that was pretrained by ImageNet and completed the transfer learning for classification, and finally, the result of the classification was obtained. Next, we elaborate on each part of the model in Fig. 2.

3.1 Taming transformers

Taming Transformers [13] are used to generate a variety of high-quality images; in contrast to CNNs, they contain no inductive bias that prioritizes local interactions. It uses a convolutional VQGAN to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong
compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

### 3.2 CCG-taming transformers

An image \( x \in \mathbb{R}^{H \times W \times 3} \) is represented by the spatial collection of codebook entries as \( z_q \in \mathbb{R}^{h \times w \times n_z} \), and \( n_z \) represents the dimensionality of codes. First learn a convolutional network composed of an encoder \( E \) and a decoder \( G \), and it will learn the code to represent the image discrete codebook \( Z = \{z_k\}_{k=1}^K \subset \mathbb{R}^{n_z} \). Image \( x \) passes through encoder \( E \), We obtain \( z_q \) using the encoding \( \hat{z} = E(x) \in \mathbb{R}^{h \times w \times n_z} \) and a subsequent elementwise quantization \( q(\cdot) \) of each spatial code \( \hat{z}_{ij} \in \mathbb{R}^{n_z} \) onto its closest codebook entry \( z_k \):

\[
z_q = q(\hat{z}) = \left( \arg\min_{z_k \in Z} \| \hat{z}_{ij} - z_k \| \right) \in \mathbb{R}^{h \times w \times n_z} \\
\hat{z} = G(z_q) = G(q(E(x)))
\]

\[
L_{VQ} = (E, G, Z) = \|x - \hat{x}\|^2 + \|\text{sg}(E(x) - z_q)\|^2 + \beta \|\text{sg}(z_q - E(x))\|^2
\]
Where \( \| x - \hat{x} \|_2^2 \) is a reconstruction loss, \( \text{sg}[\cdot] \) denotes the stop-gradient operation, and \( \| \text{sg}[z_q] - E(x) \|_2^2 \) is the so-called “commitment loss” with weighting factor \( \beta \).

VQGAN uses a discriminator and perceptual loss to maintain good perceptual quality at an increased compression rate. VQGAN replaces B in A with a discriminator D to distinguish between real and reconstructed images.

\[
L_{\text{GAN}}(\{E, G, Z\}, D) = [\log D(x) + \log(1-D(\hat{x}))]
\]

The complete objective for finding the optimal compression model \( Q^* \) = \{E*, G*, Z*\} then reads:

\[
Q^* = \underset{E, G, Z}{\text{argmin}} \max_D \mathbb{E}_{x \sim p(x)}[L_{\text{VQ}}(E, G, Z) + \lambda L_{\text{GAN}}(\{E, G, Z\}, D)]
\]

Where we compute the adaptive weight \( \lambda \) according to:

\[
\lambda = \frac{\nabla_{G_L} L_{\text{rec}}}{\nabla_{G_L} L_{\text{GAN}}} + \delta
\]

Where \( L_{\text{rec}} \) is the perceptual reconstruction loss [71], \( \nabla_{G_L}[\cdot] \) denotes the gradient of its input w.r.t. the last layer \( L \) of the decoder, and \( \delta = 10^{-6} \) is used for numerical stability. With \( E \) and \( G \) available, the cervical cell images can be represented in terms of the codebook indices of their encodings. The quantized encoding of an image \( x \) is given by \( z_q = q(E(x)) \in \mathbb{Z}^{h \times w \times n_z} \) and is equivalent to a sequence \( s \in \{0, \ldots, \vert Z \vert - 1\}^{h \times w} \) of indices from the codebook, which is obtained by replacing each code by its index in the codebook \( Z \):

\[
s_{ij} = k \text{such that } z_{q,i,j} = z_k
\]

By mapping indices of a sequence \( s \) back to their corresponding codebook entries, \( z_q = (z_{q,s}) \) is readily recovered and decoded to an image \( \hat{x} = G(z_q) \). Thus, after choosing some ordering of the indices in \( s \), image generation can be formulated as autoregressive next-index prediction: given indices \( s_{<i} \), the transformer learns to predict the distribution of possible next indices, i.e., \( p(s_i \vert s_{<i}) \) to compute the likelihood of the full representation as \( p(s) = \prod_i p(s_i \vert s_{<i}) \). This allows us to directly maximize the log-likelihood of the data representations (As shown in equation (8)), and Fig. 3 is the overview of the framework for the CCG-taming transformers.

\[
L_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)}[-\log p(s)]
\]

### 3.2.1 SE-block and MultiRes-block

Figure 4 shows the structure of SE-block and MultiRes-block. We add the SE (squeeze-and-stimulate) block [29] in encoder \( E \) of VQGAN, which can adaptively recalibrate the characteristic response of each channel by explicitly modeling the interdependence between channels. The SE-block has excellent generalization ability in different datasets and brings significant performance improvement to the convolutional neural network. To improve the image information learning ability of VQGAN for cervical cancer Pap cell smears, this paper introduces the convolutional block attention module (CBAM) [67] to enhance VQGAN’s attention to cells and nuclei.
In Fig. 4, the dimension information in the excitation block represents the output of the layer. First, the global average-pooling layer in the squeeze block is used for compression, and a bottleneck structure is formed by two fully connected layers to construct the correlation between channels and export the same weight as the input feature. After that, the feature dimension is reduced to 1/16 of the input, and then the initial feature dimension is restored through the full connection layer after the activation of ReLU. Then, a weight between 0 and 1 is obtained through the sigmoid activation function, and the weight is weighted to the features of different feature channels through the scale operation. Finally, the feature of the residual

Fig. 3 Overview of the framework for the CCG-taming transformers

Fig. 4 Structure diagram of the encoder and decoder in the CCG-taming transformers
block is redefined before the addition. The structure of the decoder is the same as the structure of the encoder, the difference is that the spatial structure is opposite to the encoder structure, and down-sampling becomes up-sampling. The specific structure is shown in Fig. 5.

The introduction of the SE-block makes the classification model of cervical cancer cell images more nonlinear, which can better fit the complex correlation between channels, greatly reduce the number of parameters and calculations, and solve the problem of gradient disappearance near the input layer on the trunk, which makes the model easy to optimize and increases the generalization ability of the classification model.

We improved the convolution structure in the encoder and introduced the MultiRes-block [25], which was originally used in the image segmentation model decoder to extract image information. The MultiRes-block is introduced to improve the efficiency of extracting cervical cancer cell images and optimize the structure of the decoder to replace the redundant structure of multilayer convolution and superposition in the original decoder.

We also add a residual connection because of their efficacy in medical image information extraction and for the introduction of the $1 \times 1$ convolutional layers, which may allow us to comprehend some additional spatial information. We call this arrangement a MultiRes-block, as shown in Fig. 5.

Figure 5 Illustrates the MultiRes-block, where we increased the number of filters in the successive three layers gradually and added a residual connection (along with a $1 \times 1$ filter for conserving dimensions). Res paths of MultiRes-blocks introduce some additional processing to make the two feature maps more homogeneous

3.2.2 Layer normalization

The input of neurons in the same layer in LN has the same mean and variance, and different input samples have different mean and variance; therefore, LN does not depend on the size of the batch and the depth of the input sequence, and is used for a new word representation of Attention output. For standardization processing, adding Layer Normalization processing is to standardize the data, which is convenient for the subsequent non-linear processing of the data by the ReLU activation function in Feed Forward. Normalize the data to the active area of the ReLU activation function through Normalization, which can make the activation function work better.

LN is the normalize operation for the input of all neurons in a certain layer of the deep network according to the following formula.

\[
\text{Input} \rightarrow \text{Conv} \ 3 \times 3 \rightarrow \text{Conv} \ 3 \times 3 \rightarrow \text{Conv} \ 3 \times 3 \rightarrow \text{Concatenation} \rightarrow \text{Output}
\]

Fig. 5 Structure diagram of MultiRes-block

Springer
For the $i^{th}$ summed input in the $h^{th}$ layer, $H$ denotes the number of hidden units in a layer. $a_i$ is the vector representation of the summed inputs to the neurons in layer $i^{th}$ all the hidden units in a layer share the same normalization terms $\sigma$.

3.2.3 SMOTE-Tomek links

To balance the source data set and the number of samples and weights of the images we generate, we used Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links Under sampling in a pipelined approach [50]. Due to random oversampling, the strategy of simply copying samples is adopted to increase the minority samples. This is easy to cause the problem of model overfitting, that is, the information learned by the model is too special and not general enough. The basic idea of the SMOTE algorithm is to analyze the minority samples and artificially synthesize new samples based on the minority samples and add them to the data set. There is a parameter that represents the percentage of oversampling, and its value represents the number of synthetic samples to be created. For each minority instance, the $k$ nearest neighbors make them belong to the same class.

$$k = \frac{(\text{SMOTE}\%)}{100}$$

In this case, the minority class is oversampled with the applied ‘sampling-strategy’ parameter represented as ‘$k$’ ($k = 0.5$) i.e. keeping the sampling strategy parameter as 0.5 increases the number of minority class examples by 50%.

However, SMOTE has some blindness when selecting neighbors. Therefore, we use Tomek Links method for under-sampling to clean up noisy samples. Tomek Links can find more samples of opposite categories and delete the majority of samples in the pair. After such processing, the dividing line between the melanoma and benign has become clearer, making the existence of the minority more obvious.

A Tomek Link is the distance between two samples from two different classes say $x$ and $y$ such that for any sample $z$:

$$d(x, y) < d(x, z) \text{ and } d(x, y) < d(y, z)$$

In a pipelined approach the minority Class is oversampled by using SMOTE followed by removing the majority class samples by Tomek Links.

3.3 Tokens-to-token vision transformers (T2T-ViT)

Tokens-to-token vision transformers (T2T-ViT) [70] can progressively tokenize an image to tokens and have an efficient backbone. T2T-ViT consists of two main components (Fig. 6): the tokens-to-token module (T2T module) is used to model the local structure information of images and reduce the token length progressively; the T2T-ViT backbone is used to draw the global attention relations on the tokens from the T2T module.
3.3.1 T2T-ViT structure

Each T2T process has two steps: restructurization and soft splitting. The T2T transformer is shown in Fig. 7:

![Diagram of T2T-ViT structure](image)

**Fig. 6** The overall network architecture of T2T-ViT. The input cervical cell image is first soft split as patches, and then unfolded as a sequence of tokens $T_i$. The length of tokens are reduced gradually in the T2T, after two iterations output $T_f$. Then, the backbone takes the fixed tokens as input and output the predictions, PE is position embedding.

**Fig. 7** The tokens $T_i$ are restructurized as image $I_i$ after a transformation and reshaping, then $I_i$ is split with overlapping to tokens $T_{i+1}$ again. Tokens-to-Token in the T2T process: the four tokens (1,2,4,5) of the input $I_i$ are concatenated to form one token in $T_{i+1}$.
The restructurization is shown in Fig. 7, giving a sequence of tokens $T$ from the last layer; here, $T$ denotes tokens from the last layer. $T$ is transformed by the self-attention block.

$$T' = \text{MLP}(\text{MSA}(T))$$

(12)

Where MSA denotes the multihead self-attention operation with layer normalization and MLP is the multilayer perceptron with layer normalization in the standard transformer [11]. Then, the tokens are reshaped as images on the spatial dimension.

$$I = \text{Reshape}(T')$$

(13)

Where “Reshape” reorganizes tokens $T' \in \mathbb{R}^{l \times c}$ to $I \in \mathbb{R}^{h \times w \times c}$, where $l$ is the length of $T'$, $h$, $w$, and $c$ are the height, width and channel, respectively, and $l = h \times w$.

As shown in Fig. 7, after obtaining the restructurized image $I$, a soft split is applied to model the local structure information and reduce the token length. Specifically, to avoid information loss in generating tokens from the restructured image, we split the cervical cell image into patches with overlap. As such, each patch is correlated with surrounding patches to establish prior knowledge that there should be stronger correlations between surrounding tokens. The tokens in each split patch are concatenated as one token (tokens-to-token, Fig. 7); thus, the local information can be aggregated from surrounding pixels and patches. When conducting the soft split, the size of each patch is $k \times k$ with $s$ overlapping and $p$ padding on the image, where $k - s$ is similar to the stride in the convolution operation. Therefore, for the reconstructed image $I \in \mathbb{R}^{h \times w \times c}$, the length of output tokens $T_O$ after soft splitting is

$$l_O = \left\lceil \frac{h + 2p - k}{k - s} + 1 \right\rceil \times \left\lceil \frac{w + 2p - k}{k - s} + 1 \right\rceil$$

(14)

Each split patch has size $k \times k \times c$. We flatten all patches in spatial dimensions to tokens $T_O \in \mathbb{R}^{l_0 \times c_0 \times k^2}$. After the soft split, the output tokens are fed to the next T2T process.

### 3.3.2 T2T-ViT backbone and transfer learning

The T2T-ViT has two parts: the tokens-to-token (T2T) module and the T2T-ViT backbone (Fig. 6). There are various possible design choices for the T2T module. Here, we set $n = 2$ as shown in Fig. 8, which means there are $n + 1 = 3$ soft split and $n = 2$ restructurization in the

![Fig. 8](image)

(a) Images generated by depth-guided neural rendering  
(b) Samples generated from semantic layouts
The T2T module. The patch size set for the three soft splits is $P = [7, 3, 3]$, and the overlapping set is $S = [3, 1, 1]$, which reduces the size of the input image from $224 \times 224$ to $14 \times 14$ according to Eq. 14.

By conducting the above restructuring and soft splitting iteratively, the T2T module can progressively reduce the token length and transform the spatial structure of the image. The iterative process in the T2T module can be formulated as:

$$T'_i = \text{MLP}(\text{MSA}(T_i))$$

$$I_i = \text{Reshape}(T'_i)$$

$$T_{i+1} = \text{SS}(I_i), i = 1 \ldots (n-1)$$

For the input cervical cell image $I_0$, we apply a soft split first to split the image to tokens: $T_1 = \text{SS}(I_0)$. After the final iteration, the output tokens $T_f$ of the T2T module have a fixed length; thus, the backbone of T2T-ViT can model the global relations on $T_f$. In this paper, the backbone of T2T-ViT selects the T2T-ViT-ResNeXt with the highest classification accuracy of the original T2T-ViT.

The mathematical formulation of a classifier with a training algorithm is shown below:

| Algorithm 1: Training the Classification Model Based on CCG-Taming Transformers and T2T-ViT |
|-----------------------------------------------|
| **Input:**                                     |
| 1: The imaged of ImageNet, $S_{\text{lm}}$;    |
| 2: The cervical cancer cell image of cervical cell datasets, $S_{\text{cen}}$; |
| Use $S_{\text{lm}}$ to operate Transfer Learning in Classification Model Based on CCG-Taming Transformers and T2T-ViT, $M' = \text{TL}(M(S_{\text{lm}}))$; |
| Preprocessing of $S_{\text{cen}}$: $\text{CCG-Taming Transformers(})$ is the operation of generating samples using CCG-Taming Transformers, and $\text{SMOTE(})$ is the operation of SMOTE-Tomek Links on samples; |
| $S'_{\text{cen}} = \text{SMOTE}(\text{CCG-Taming Transformers}(S_{\text{cen}}))$; |
| Train the preprocessed $S'_{\text{cen}}$ by using $M'$, which has been transferred and learned, $N$ is the class number of dataset; |
| $\sum_{i=1}^{N} S_i = M'(S'_{\text{cen}})$; |
| **Output:**                                    |
| 1: different sets of cervical cancer cells classified by $M'$, $\sum_{i=1}^{N} S_i$; |
| 2: Confusion Matrix of Classification; diagram of Accuracy and Loss; |
| 3: return ACC, SE, SP, H-mean, $F_1$; |

\[ \sum_{i=1}^{N} S_i \]
4 Experiment

4.1 Dataset

To evaluate the classification performance of our method, three public cervical cell image datasets are used in this paper:

The liquid-based cytology Pap smear dataset [24] consists of a total of 963 LBC images subdivided into four sets representing the four classes: NILM, LSIL, HSIL, and SCC. It comprises precancerous and cancerous lesions related to cervical cancer as per standards under the Bethesda System (TBS).

SIPaKMeD [43] consists of 4049 images of isolated cells with 5 different categories from 966 cluster cell images of Pap smear slides. The dataset comprises cells belonging to five categories: (1) dyskeratotic, (2) koilocytotic, (3) metaplastic, (4) parabasal and (5) superficial-intermediate. Classes 1–2 represent abnormal cervical cells, classes 4–5 represent normal cervical cells, and class 3 represents benign cells.

The Herlev dataset [27] consists of 917 isolated single-cell images, that is, the images contain one cervical cell. There are a total of seven classes: (1) superficial squamous epithelia, (2) superficial squamous epithelia, (3) columnar epithelia, (4) mild squamous non-keratinizing dysplasia, (5) moderate squamous nonkeratinizing dysplasia, (6) severe squamous nonkeratinizing dysplasia and (7) squamous cell carcinoma in situ intermediate. Classes 1–3 are normal cervical cells, whereas classes 4–7 are abnormal cervical cells. Tables 1, 2, 3 lists the data distribution of various samples of the three cervical cancer datasets liquid-based cytology Pap smear, SIPaKMeD, and Herlev.

4.2 Model training process

To train the cervical cancer cell classification model, the weights of our T2T-ViT were initialized by the transfer parameters from the pretrained T2T-ViT on the ImageNet dataset. All training data were first resized to 224 × 224 × 3 and divided into minibatches for training. Transfer learning is a machine learning method that uses existing knowledge to solve different but related domain problems [55]. We used the ImageNet dataset to pretrain the V2T-ViT. Finally, we trained the cervical cancer cell images through V2T-ViT pretrained by ImageNet. During training, we set the minibatch size to 32. We used AdamW [37] as the optimizer and cosine learning rate decay [36]. The experiments were run on Python 3.7.1 in a Windows operating system with an i9-9900k processor, 64 GB memory, and one GTX2080Ti graphics card. The entire experiment was based on the open-source deep learning framework PyTorch 1.0 in Anaconda3. We trained T2T-ViT for 100 epochs, 5 training-test folds cross-validation was used for the evaluation of classification performance, our experiment configurations refers

| Category | Quantity | Property |
|----------|----------|----------|
| NILM     | 613      | normal   |
| LSIL     | 163      | abnormal |
| HSIL     | 113      |          |
| SCC      | 74       |          |
| Total image | 963  |          |
to the setting of [15, 51], we use 80% of the data in iteration. The remaining 20% is used for testing. Specifically, 4 of 5 folds were used as the training set and the other as the validation set for 5 rounds. The classification evaluation metrics were obtained by averaging the results from the 5 validation sets. After five iterations, all the data were tested.

The hyperparameters maintained during training are shown in Table 4.

### 4.3 Evaluation metrics

#### 4.3.1 Evaluation metrics of cervical cell classification

To comprehensively evaluate the classification performance of the model, accuracy (ACC), sensitivity (SE), specificity (SP), H-mean, and F1 score were used as evaluation metrics. Accuracy is the overall percentage of correctly identified cells and can be used to evaluate the ability of the classifier to judge an overall sample. Sensitivity is also called the true positive rate or recall, which reports the proportion of correctly identified abnormal cells, specificity reports the proportion of correctly identified normal cells, F1 score is the harmonic mean of precision and recall.

\[
ACC = \frac{N_{TP} + N_{TN}}{N_{FP} + N_{TP} + N_{TN} + N_{FN}} \tag{18}
\]

\[
SE(\text{recall}) = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{19}
\]
\[ SP = \frac{N_{TN}}{N_{TN} + N_{FP}} \]  

\[ H\text{-}mean = 2 \times \frac{SE \times SP}{SE + SP} \]  

\[ F_1 = \frac{2SE \cdot N_{TP}}{N_{TP} + N_{FP}} \]  

\[ IS(G) = \exp\left( \mathbb{E}_{x \sim p_x} D_{KL}(p(y|x) \| p(y)) \right) \]  

\[ D_{KL}(p(y|x), p(y), y, x \text{ denote the KL divergence formula, the conditional label distribution of samples, the marginal distribution obtained from all the samples, the given picture, and the main object in the picture, respectively. A high IS indicates that the sample image is similar to a particular ImageNet category.} \]

The Fréchet inception distance (FID) [23] embeds a set of generated images into a feature space represented by a specific layer of inception or any CNN [47]. It uses a continuous multivariate Gaussian distribution to represent the embedding feature distributions of the real data and the generated data, and the Fréchet distance between these two Gaussian distributions is computed by:

\[ \text{FID}(r, g) = \left\| \mu_r - \mu_g \right\|_2^2 + \text{Tr} \left( C_r + C_g - 2(C_r C_g)^{\frac{1}{2}} \right) \]  

\( N_{TP}, N_{TN}, N_{FP}, N_{FN} \) denote the number of true positive, true negative, false positive, false negative, respectively.

### 4.3.2 Evaluation metrics of cervical cell generation

The inception score (IS) [49], which is a measure of the generated image quality, uses a pretrained inception network [47] to extract the features of generated images by computing:

\[ IS(G) = \exp\left( \mathbb{E}_{x \sim p_x} D_{KL}(p(y|x) \| p(y)) \right) \]  

\[ D_{KL}(\cdot, \cdot), p(y|x), p(y), y, \text{ and } x \text{ denote the KL divergence formula, the conditional label distribution of samples, the marginal distribution obtained from all the samples, the given picture, and the main object in the picture, respectively. A high IS indicates that the sample image is similar to a particular ImageNet category.} \]

The Fréchet inception distance (FID) [23] embeds a set of generated images into a feature space represented by a specific layer of inception or any CNN [47]. It uses a continuous multivariate Gaussian distribution to represent the embedding feature distributions of the real data and the generated data, and the Fréchet distance between these two Gaussian distributions is computed by:

\[ \text{FID}(r, g) = \left\| \mu_r - \mu_g \right\|_2^2 + \text{Tr} \left( C_r + C_g - 2(C_r C_g)^{\frac{1}{2}} \right) \]  

\( \mu_r \) and \( C_r \) denote the mean and covariance of the real data, and \( \mu_g \) and \( C_g \) denote the mean and covariance of the generated data, respectively. FID serves as a good measure for GANs due to its good discriminability, robustness, and computational efficiency.
4.4 Results

4.4.1 CCG-taming transformers generated results

We used data from each type of cervical cancer cell in three datasets (liquid-based cytology Pap smear, SIPaKMeD, and Herlev) to generate more high-quality cervical cell images and included them in the classification experiments. In this paper, the international common evaluation IS (inception score) is used to evaluate the quality of the generated cervical cancer cell image. IS is one of the most commonly used methods for evaluating images generated by GANs. The higher the value is, the more realistic the image, and the higher the quality. IS focuses on two aspects of image performance: one is the clarity of the generated image, and the other is the diversity of the generated image.

We also use the Fréchet inception distance (FID) as another cervical cell generation evaluation metric; the FID score ranges from 0 to positive infinity, and a smaller score indicates a better model. It embeds a set of generated images into a feature space represented by a specific layer of inception or any CNN [47]. The specific evaluation metrics of cervical cell generation are shown in Table 5. The comparison results with other generation models are shown in Table 11.

Figure 8a shows the images generated by depth-guided neural rendering on RIN with \( f = 16 \) using the sliding attention window. Figure 8b shows the images generated from semantic layouts on S-FLCKR with \( f = 16 \) using the sliding attention window. The generated images based on three different cervical cancer cell datasets (liquid-based cytology Pap smear dataset, SIPaKMeD, Herlev) are shown in Figs. 9, 10, 11. Through the CCG-taming transformers, we invited professional inspection experts to distinguish them, and the results showed that the generated cervical cancer cell images have little difference from the real images, which can thus be used as the training dataset.

This paper balances the weights and quantity of different classes and expands the number of various samples on three different cervical cancer cell datasets (liquid-based cytology Pap smear dataset, SIPaKMeD, and Herlev) by SMOTE-Tomek Links. We named these three new datasets CCG1, CCG2, and CCG3. The distribution of different classes of cervical cells is shown in Tables 6, 7, 8.

4.4.2 Cervical cell classification results

In this section, we describe the classification results on three cervical cancer cell datasets: (1) the liquid-based cytology Pap smear dataset, (2) SIPaKMeD and (3) Herlev. After 150 epochs of classification in dataset (1), the whole model yielded good results, the test loss reached a minimum, and the corresponding test accuracy of dataset (1) was 98.79%. After 100 epochs of classification in datasets (2) and (3), the model’s curve fluctuations tended to be stable, the test loss reached a minimum, and the corresponding test accuracies of datasets (2) and (3) were

| Method                      | IS   | FID   | Recall | Precision |
|-----------------------------|------|-------|--------|-----------|
| CCG-Taming Transformers     | 3.75 | 0.71  | 0.32   | 0.65      |
99.58% and 99.98%, respectively. Tables 9, 10 shows the classification ACC, SE, SP, H-mean, and F1 score for datasets (1–3) and CCG1–3. The classification results of T2T-ViT in three different datasets are significantly better than those of the other classification models. Although ViT also achieved good classification accuracy, it was not superior to all classification models based on CNN.

Fig. 9 Cervical cancer cell images generated by CCG-taming transformers based on the liquid-based cytology Pap smear dataset

Fig. 10 Cervical cancer cell images generated by CCG-taming transformers based on the SIPAKMeD dataset

Fig. 11 The cervical cancer cell images generated by CCG-taming transformers based on the Herlev dataset
Figures 12, 13, 14a shows the training and test accuracy versus the number of training epochs in the (1–3) dataset, and Figs. 12, 13, 14b shows the training and test loss versus the number of training epochs in the CCG-1-3 dataset.

A confusion matrix, also known as an error matrix, is a standard format for accuracy evaluation. The cervical cancer cell classification model output is divided into different categories \( n \), so it is represented by an \( n \times n \) matrix. For the dataset (1–3), \( n \) is equal to 4, 5, and 7. For the CCG1–3 dataset, \( n \) is equal to 4, 5, and 7 (Fig. 15).

5 Discussion

Cervical cell classification is a challenging issue for automatic screening of cervical cytology. The performance of the classification method determines whether it can bring convenience to cytoscreeners and women. As the current mainstream classification framework, a convolutional neural network (CNN) uses a hierarchical deep architecture to automatically learn high-level features.

For cervical cancer Pap cell smears, there are only two public datasets SIPaKMeD and Herlev in previous research, and the dataset has a small sample size. The number and weight of various samples are unbalanced, and the sample image quality is uneven. This limits future research on cervical cancer and the timeliness of clinical analysis of the patient’s condition. To solve these problems, this paper proposes CCG-taming transformers and generates sample images based on the existing published cervical cancer Pap cell smear dataset.

We combine CNN with the transformer, use the convolution method to efficiently learn the codebook of the context-rich visual part, and then learn its global composition model, use CNN-based VQGAN to learn the codebook of the context-rich visual part, and the components of the codebook follow the autoregressive transformer architecture to model. A discrete codebook provides the interface between these architectures, and the patch-based discriminator achieves powerful compression capabilities while maintaining high perceptual quality.

To better verify the experimental results of this paper, we used CCG-taming transformers, StyleGAN [31], PGAN [30], DC-GAN [45], LAPGAN [9] and GAN [16] to generate 1000 cervical cell images, and the samples were evaluated 20 times. The experimental results are shown in Table 11 and Figs. 16, 17, 18.

| Table 6 | Data distribution of the CCG1 dataset |
|---------|--------------------------------------|
| Category | NILM | LSIL | HSIL | SCC | Total quantity |
| Original quantity | 613 | 163 | 113 | 74 | 963 |
| Added quantity | 387 | 837 | 887 | 926 | 3037 |
| Total quantity | 1000 | 1000 | 1000 | 1000 | 4000 |

| Table 7 | Data distribution of the CCG2 dataset |
|---------|--------------------------------------|
| Category | Dyskeratotic | Koilocytic | Parabasal | Metaplastic | Superficial/Intermediate | Total quantity |
| Original quantity | 813 | 787 | 825 | 813 | 793 | 4049 |
| Added quantity | 687 | 713 | 675 | 687 | 707 | 3451 |
| Total quantity | 1500 | 1500 | 1500 | 1500 | 1500 | 7500 |
| Category | Superficial squamous epithelial | Intermediate squamous epithelial | Columnar epithelial | Mild squamous non-keratinizing dysplasia | Moderate squamous nonkeratinizing dysplasia | Severe squamous nonkeratinizing dysplasia | Squamous cell carcinoma in situ | Total quantity |
|----------|--------------------------------|---------------------------------|--------------------|----------------------------------------|-------------------------------------------|----------------------------------------|-------------------------------|----------------|
| Original quantity | 74 | 70 | 98 | 182 | 146 | 197 | 150 | 917 |
| Added quantity | 926 | 930 | 902 | 818 | 854 | 803 | 850 | 6083 |
| Total quantity | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 7000 |
According to the above table, the IS value of the proposed CCG-taming transformers is the highest, reaching 3.759. The IS values (the higher the better) of the other five methods are 0.705, 1.273, 1.235, 1.431, 1.791 lower, the FID value (the lower the better) of the proposed CCG-taming transformers is the lowest, reaching 0.714, and the FID values of the other four methods are 1.414, 0.854, 0.643, 1.303 and 2.533 lower. The values of precision and recall are all from 0 to 1, and values close to 1 indicate good results. The CCG-taming transformers

| Dataset                  | Method      | ACC (%)     | SE (%)     | SP (%)     | H-mean (%)   | F1-score (%) |
|--------------------------|-------------|-------------|------------|------------|--------------|--------------|
| Liquid-based cytology    | AlexNet     | 86.67±0.15  | 97.55±0.54 | 88.24±2.89 | 93.62±0.84  | 87.08±1.72   |
|                          | ResNet50    | 90.97±0.57  | 97.99±0.35 | 95.97±0.71 | 96.46±1.57  | 90.98±3.41   |
|                          | DenseNet201 | 92.99±0.32  | 94.98±0.47 | 90.84±2.78 | 96.39±1.27  | 89.99±0.95   |
|                          | ShuffleNetV2| 95.33±0.85  | 96.94±1.24 | 91.14±0.57 | 90.31±0.89  | 90.73±0.47   |
|                          | ViT         | 96.48±1.24  | 95.01±0.57 | 92.54±1.57 | 92.47±0.98  | 92.78±1.25   |
|                          | T2T-ViT-24  | 98.79±0.82  | 98.61±0.62 | 97.54±0.71 | 98.47±0.56  | 99.19±0.46   |
| SIPaKMeD                 | AlexNet     | 88.71±0.56  | 96.52±0.82 | 89.17±1.58 | 91.21±1.28  | 90.87±1.24   |
|                          | ResNet50    | 93.87±0.75  | 95.48±0.75 | 90.76±0.85 | 95.47±0.78  | 97.48±0.68   |
|                          | DenseNet201 | 95.99±0.58  | 95.47±1.25 | 92.48±0.62 | 91.46±0.87  | 94.41±0.37   |
|                          | ShuffleNetV2| 97.79±1.75  | 94.15±2.57 | 99.45±0.24 | 96.81±0.27  | 97.15±0.85   |
|                          | ViT         | 96.57±1.85  | 98.84±0.42 | 92.57±1.87 | 97.84±1.27  | 95.28±0.51   |
|                          | T2T-ViT-24  | 99.58±0.25  | 98.34±0.25 | 98.57±0.72 | 98.78±0.85  | 99.75±0.08   |
| Herlev                   | AlexNet     | 90.41±1.27  | 95.47±0.64 | 90.72±0.95 | 95.48±0.28  | 90.51±0.76   |
|                          | ResNet50    | 95.54±0.28  | 98.54±0.74 | 92.58±0.38 | 95.41±0.84  | 94.21±0.21   |
|                          | DenseNet201 | 96.87±0.98  | 96.52±0.83 | 91.57±0.74 | 95.51±0.57  | 92.47±0.67   |
|                          | ShuffleNetV2| 98.54±0.58  | 96.14±1.57 | 97.25±2.54 | 98.57±0.28  | 97.54±0.09   |
|                          | ViT         | 97.27±2.57  | 97.29±1.28 | 95.64±1.87 | 96.84±2.87  | 95.75±0.27   |
|                          | T2T-ViT-24  | 99.98±0.01  | 99.87±0.05 | 99.64±0.15 | 98.85±0.97  | 99.86±0.05   |

According to the above table, the IS value of the proposed CCG-taming transformers is the highest, reaching 3.759. The IS values (the higher the better) of the other five methods are 0.705, 1.273, 1.235, 1.431, 1.791 lower, the FID value (the lower the better) of the proposed CCG-taming transformers is the lowest, reaching 0.714, and the FID values of the other four methods are 1.414, 0.854, 0.643, 1.303 and 2.533 lower. The values of precision and recall are all from 0 to 1, and values close to 1 indicate good results. The CCG-taming transformers

| Dataset                  | Method      | ACC (%)     | SE (%)     | SP (%)     | H-mean (%)   | F1-score (%) |
|--------------------------|-------------|-------------|------------|------------|--------------|--------------|
| CCG1                     | ResNet101   | 87.85±2.82  | 92.54±0.85 | 90.71±0.71 | 91.57±0.38  | 90.24±0.71   |
|                          | MobileNetV2 | 89.15±0.81  | 94.78±0.28 | 98.33±0.34 | 97.42±0.51  | 94.32±0.14   |
|                          | ViT         | 92.47±1.08  | 96.84±1.21 | 96.81±1.27 | 96.72±0.84  | 95.57±0.53   |
|                          | T2T-ViT-24  | 94.18±0.24  | 97.84±0.72 | 98.42±0.73 | 94.37±0.53  | 98.14±1.02   |
| CCG2                     | ResNet101   | 90.54±0.51  | 96.84±0.37 | 95.27±0.61 | 97.51±0.27  | 96.57±0.34   |
|                          | MobileNetV2 | 92.82±1.11  | 95.87±0.28 | 98.17±0.11 | 95.47±1.76  | 97.14±0.11   |
|                          | ViT         | 94.58±1.54  | 97.82±0.15 | 97.57±0.21 | 96.48±0.74  | 97.57±1.72   |
|                          | T2T-ViT-24  | 96.77±0.25  | 96.13±0.47 | 97.64±0.92 | 98.55±0.47  | 96.71±1.25   |
| CCG3                     | ResNet101   | 92.41±1.52  | 93.47±2.48 | 90.57±0.85 | 96.57±2.17  | 93.27±1.57   |
|                          | MobileNetV2 | 95.67±0.37  | 96.81±0.42 | 94.81±0.51 | 98.55±0.13  | 94.28±0.18   |
|                          | ViT         | 93.34±0.37  | 93.54±0.19 | 92.38±0.27 | 93.42±0.33  | 91.57±0.42   |
|                          | T2T-ViT-24  | 97.86±0.15  | 98.57±1.02 | 97.71±1.31 | 98.87±0.25  | 98.52±0.74   |
Fig. 12  Accuracy and loss diagram of liquid-based cytology Pap smear dataset classification. a Training and test loss versus the number of training epochs, b training and test BMA training epochs

Fig. 13  Accuracy and loss diagram of SIPaKMeD. a Training and test loss versus the number of training epochs, b training and test BMA training epochs

Fig. 14  Accuracy and loss diagram of Herlev. a Training and test loss versus the number of training epochs, b training and test BMA training epochs
Fig. 15 Confusion matrix of T2T-ViT-24 classification results on different datasets. a Liquid-based cytology Pap smear dataset, b SIPaKMeD, c Herlev, d CCG1, e CCG2, and f CCG3

Table 11 The average generation metrics on three datasets: (a) liquid-based cytology Pap smear dataset, (b) SIPaKMeD and (c) Herlev

| Method          | CCG-taming transformers | StyleGAN | PGAN | DC-GAN | LAPGAN | GAN    |
|-----------------|-------------------------|----------|------|--------|--------|--------|
| IS              | 3.75                    | 3.05     | 2.48 | 2.52   | 2.32   | 1.96   |
| FID             | 0.71                    | 2.12     | 1.56 | 1.35   | 2.01   | 3.24   |
| Recall          | 0.32                    | 0.20     | 0.26 | 0.14   | 0.12   | 0.08   |
| Precision       | 0.65                    | 0.28     | 0.45 | 0.58   | 0.47   | 0.29   |
scores best in four metrics compared to the other four models, and it is proven to be a better image synthesis model comprehensively, which indicates that the image quality generated by the CCG-taming transformers used in this paper is higher.

It can be seen in Figs. 16, 17, 18 that images of cervical cancer cells generated based on various types of GANs have many problems. For each type of GAN, the superposition of different image styles may cause image distortion or image components of the same style (Figs. 16, 17). This is because CNN-based GAN processes images according to pixels. Different styles tend to be overly rigid when fused or simply repeat directly; the generated sample images may appear blurred (Fig. 18, row 2) because GAN’s CNN has a limited depth of structure. In the case of low resolution and general quality of the original dataset image, the GAN’s identification area will have a low standard for judging the generated samples. These low-quality generated cervical cell images will undoubtedly affect the classification model and also increase the distinguishing difficulty in clinical diagnosis by gynecologists.

The CCG-taming transformers serialize the image to analyze the image information. Through the information interaction between VQGAN and the codebook, more deep details of the image can be learned, making the transition between cervical cancer cells in the generated image more naturally, it avoids the problems that may occur in the above-generated images, and makes the cervical cancer sample images generated by CCG-taming transformers closer to the real images to provide more objective reference and valuable sample images for cervical cancer research, which should assist gynecologists in clinical diagnosis and improve the survival rate of cervical cancer patients.

We conducted ablation experiments on the innovative part (SMOTE-Tomek Links, MultiRes-block, SE-block and Layer normalization) of this paper to verify the effectiveness of the innovative part of this paper for improving the performance of the experiment. The results, averaged over the three datasets, are shown in Table 12.

![Fig. 16 Abnormal images generated by various types of GANs on the SIPAKMeD dataset](image1)

![Fig. 17 Abnormal images generated by various types of GANs on the Herlev dataset](image2)
We first consider the use of the SMOTE-Tomek Links. It is found that the BMA with the improved skip connection is 2.9% higher. Subsequently, we consider the introduction of MultiRes-block when the SMOTE-Tomek Links is used. It is seen that the BMA with MultiRes-block is 2.4% higher. Then, assuming that the SMOTE-Tomek Links and MultiRes-block are used, we consider the introduction of the SE block. It is demonstrated that with the SE block, BMA is 2.2% higher. Finally, we consider the introduction of Layer normalization. It is seen that with Layer normalization, BMA is 1.7% higher. It is thus demonstrated that the introduction of SMOTE-Tomek Links, MultiRes-block, SE-block and Layer normalization can significantly improve the BMA of the entire segmentation model.

There have been many classification studies based on CNN’s public dataset of cervical cancer Pap cells. Tables 13, 14 compare the results of this study with other CNN-based research results.

Tables 13, 14 show that in the public cervical cancer Pap smear dataset, the classification research based on CNN achieved high classification accuracy, but this also shows that the

| SMOTE-Tomek Links | MultiRes-block | SE-block | Layer normalization | BMA(%) |
|------------------|----------------|---------|-------------------|--------|
|                  |                |         |                   | 87.8   |
|                  |                |         |                   | 90.7   |
| √                |                |         |                   | 93.1   |
| √                |                |         |                   | 95.3   |
| √                |                | √       |                   | 98.8   |

![Fig. 18 Abnormal images generated by various types of GANs on the liquid-based cytology Pap smear dataset](image)
classification of the classification research for these public datasets approached the state of overfitting. In the published cervical cancer Pap cell smear dataset, the classification model based on T2T-ViT in this paper is better than the classification accuracy based on CNN.

Win et al. [66] develop a computer-assisted screening system for cervical cancer using digital image processing of Pap smear images; Haryanto et al. [21] aims to create the classification model of Cervical Cell Images using the Convolutional Neural Network (CNN) algorithm; Talo et al. [53] used a deep learning approach to classify cervical cell images obtained from pap smear slides. Mamunur et al. [38] proposed DeepCervix, a hybrid deep feature fusion (HDFF) technique based on DL to classify the cervical cells accurately; Dounias et al. [12] compared the performance of various intelligent methodologies in the task of pap-smear diagnosis; Marinakis et al. [39] proposed an effective genetic algorithm scheme which is combined with a number of nearest neighbor based classifiers.; Dong et al. [10] proposed a machine learning method based on feature selection algorithm for cervical cell classification.

The above methods have achieved good results in the classification task of cervical cancer. However, the main purpose of these studies is to improve the accuracy of classification, but they did not consider the impact of the limited data set and the number of samples on the classification results. It is also one of the biggest differences between the research in this article and the above studies. In addition, there are too few public sample data sets for cervical cancer, which leads to serious overfitting of cervical cancer classification studies at this stage, which seriously hinders cervical cancer research. In this paper, the accuracy of classification is improved by balancing the weight and number of samples between different categories, and can generate more samples of cervical cancer cells in different categories, providing more objective reference value for the research of cervical cancer; Second, most of the above studies use CNN-based networks to classify cervical cancer, and this paper combines CNN and Transformer to further improve the classification model’s ability to capture global contextual information on the feature images of the classified samples, this can also make up for the deficiencies of the CNN-based classification model and improve the performance of the classification model.

Considering that the potential relationship between cervical cell images is ignored in the process of CNN feature learning, our goal is to learn the relationship between cell images through the CBAM module in V2T-ViT to enhance the discrimination ability of V2T-ViT features.

To verify the effectiveness of balancing the sample size and weight of various types of samples and using high-quality samples for classification to improve classification accuracy, we conducted experiments on the influence of different sample types from three cervical cancer cell datasets: (1) the liquid-based cytology Pap smear dataset, (2) SIPaKMeD and (3) Table 13 Different models’ classification results in SIPaKMeD

| Year | Model(%) | Model(%) | (7-class) BMA(%) |
|------|----------|----------|-----------------|
| 2019 | ResNet-34 | GV et al. [18] | 96.37 |
| 2019 | DenseNet | Talo et al. [53] | 98.96 |
| 2020 | RF-KNN-SVM-LD | Win et al. [66] | 94.09 |
| 2020 | AlexNet-padding | Haryanto et al. [21] | 98.00 |
| 2021 | hybrid deep feature fusion (HDFF) | Mamunur et al. [38] | 99.14 |
| 2021 | Our model | – | 99.58 |
Herlev on the T2T-ViT in this paper. First, we prepared four datasets of different samples that were altered by matrix transformation: Dataset I: simply upweight the classes with few samples; Dataset II: equalize the number for each class by showing minority examples more often; Dataset III: only CCG-taming transformers generated samples shown to increase the relative weight of these classes; Dataset IV: the samples generated by CCG-taming transformers were added to the cervical cell image classes with fewer samples and then balanced with the weight of different kinds of cervical cell images; Dataset V: combined with Dataset I and Dataset II.

Then, we used the classification framework of this paper to train these four datasets and analyze the results.

According to the analysis of the experimental results obtained in Tables 15, 16, 17, we can simply overweight the classes with few samples (Dataset I) and equalize the number for each class by showing minority examples more often. Compared with the method of using CCG-taming transformers to expand fewer sample classes, Dataset II has a lower ACC and other valuating metrics, while Dataset V, formed by combining Dataset I and Dataset II, does not obtain better results. We can see that Dataset III and Dataset IV both use the method of expanding fewer sample classes through the CCG-taming transformers, and Dataset IV balances the weight of each sample class based on the Dataset III method. Finally, we find that Dataset IV requires less training time than Dataset III and obtains higher evaluation metrics.

In addition, we also conducted an image generation experiment on the ISIC2019 skin cancer dataset to verify the generalization ability of CCG-Taming Transformers. The results are shown in Table 18.

The CCG-Taming Transformers scores best in four metrics compared to the other four models, and it is proven to be a better image generation model comprehensively.
Considering that the potential relationship between cervical cell images is ignored in the process of CNN feature learning, our goal is to learn the relationship between cell images through the CBAM module in V2T-ViT to enhance the discrimination ability of V2T-ViT features. This is also the reason why the transformer-based classification model is better than CNN in the field of image classification, segmentation and pattern recognition at this stage. For the generation of medical images, various GANs, including CNNs, not only exhibit strong local deviations but also exhibit deviations from spatial invariance by using shared weights at all locations. If a more comprehensive understanding of the input is required, these deviations will become invalid. In this paper, CCG-taming transformers do not use pixels to represent the image but represent it as the synthesis of the perceptually rich image components of the codebook.

6 Applied technical analysis

When the model was tested and simulated, we found that it is necessary to combine the artificial intelligence system with the clinical experience of dermatological doctors when some indistinguishable cervical cancer cell images are found. The significance of computer-aided diagnosis is good in curative effect and prognosis. It can also assist the pathological judgment and diagnosis problems of primary dermatological clinicians. Our work can not only help gynecologists in diagnosis but also improve patient health regarding cervical cancer awareness and provide a valuable reference for cervical cancer cell image analysis based on artificial intelligence. The research content of this paper alleviates the problem of the lack of public cervical cancer data sets and the imbalance of the weight and quantity of various samples to a certain extent. The generated images are close to the sample quality of the source data set. The images of cervical cancer cell samples we generated can save obstetricians and gynecologists.

| Method | Training time | ACC   | SE    | SP    |
|--------|--------------|-------|-------|-------|
| Dataset I | 14.5 h       | 86.42 | 91.29 | 96.78 |
| Dataset II | 17 h         | 92.81 | 96.71 | 98.27 |
| Dataset III | 18.2 h      | 97.36 | 97.27 | 98.96 |
| **Dataset IV** | **19.5 h** | **99.58** | **98.34** | **98.61** |
| Dataset V | 18.5 h       | 95.42 | 98.14 | **98.82** |

Table 16 Classification results of Datasets I-V composed of different samples in the SIPaKMeD dataset

| Method | Training time | ACC   | SE    | SP    |
|--------|--------------|-------|-------|-------|
| Dataset I | 5.5 h        | 93.54 | 90.28 | 94.83 |
| Dataset II | 6.5 h        | 92.81 | 92.37 | 95.67 |
| Dataset III | 9.2 h        | 97.57 | 99.57 | **99.71** |
| **Dataset IV** | **10.5 h** | **99.98** | **99.87** | 99.64 |
| Dataset V | 7.5 h        | 98.72 | 98.81 | 97.39 |

Table 17 Classification results of Datasets I-V composed of different samples in the Herlev dataset
from the trouble caused by low-quality samples, and improve the accuracy and efficiency of
disease analysis.

7 Limitations

Our model also has some shortcomings. There are still some technical challenges in cervical
cancer cell image classification under microscopy. Just as two kinds of cervical cancer cells
overlap in the same image, different stages of cervical cancer cells will also increase the
classification model’s difficulty, which may lead to misclassification. Although the model
shows high segmentation accuracy, these problems have not been well solved, so it is
necessary to improve the classification model further and expand the dataset samples. More
attention will focus on sample data expansion to provide better help for clinical medicine of
skin diseases in future research.

8 Conclusion

In this paper, a cervical cancer cell sample image generation model based on taming trans-
formers (CCG-taming transformers) was proposed to balance the unbalanced weights and
numbers of various class in the cervical cancer data set, and use T2T-ViT combing transfer
learning to classify the samples.

Our CCG-taming transformers improve the encoder structure by introducing and improving
MultiRes-block, SE-block and Layer normalization to improve the classification model’s
ability to capture the subtle differences in the feature map of cervical cancer cells and the
feature information of the global context; we introduce Layer Normalization to standardize the
data; we also introduce SMOTE-Tomek Links to balance the source data set and the number of
samples and weights of the images The quality of the images we generated on these three
cervical cancer data sets is very close to the source data set, the final inception score (IS),
Fréchet inception distance (FID), Recall and Precision are 3.75, 0.71, 0.32 and 0.65 respec-
tively. The proposed model is superior to other GANs based model in different quantitative
evaluation metrics.

The T2T-ViT was constructed combined with transfer learning. The classification accuracy
in the liquid-based cytology Pap smear dataset (4-class), SIPAKMeD (5-class), and Herlev (7-
class) are 98.79%, 99.58%, and 99.88%, respectively. This also proves from the side that the
image quality of cervical cancer samples generated by CCG-Taming-Transformer is better,
which promotes the improvement of the accuracy of the classification model.

The whole process of this paper, including cervical cancer cell image synthesis, classification
model construction, the balance of various sample sizes and weights, and image

| Method                    | IS   | FID  | Recall | Precision |
|---------------------------|------|------|--------|-----------|
| StyleGAN                  | 3.136| 2.245| 0.197  | 0.128     |
| DC-GAN                    | 2.216| 1.41 | 0.124  | 0.517     |
| CCG-Taming Transformers   | 3.68 | 0.96 | 0.347  | 0.681     |
classification, is applicable for other analyses of medical images, especially those datasets with intraclass-imbalanced data or insufficient labeled data. This work provides a meritorious reference for medical image analysis based on deep learning and artificial intelligence.

In future work, we will work to expand the applicability of the model in this paper, so that it can adapt to more different types of medical image data sets with a small number of samples and unbalanced weights.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11042-022-12670-0.

**Acknowledgments** Chen Zhao contributed to the writing and editing of the paper and the operation and editing of the code. Renjun Shuai (corresponding author) contributed to technological guidance and provided experimental equipment and major financial support. Li Ma contributed technical support and guidance for the paper concept. Wenjia Liu contributed to the technical guidance, and as a consultant in the medical consultant field, and Menglin Wu contributed to the direction of the paper and the funding of support and related work.

**Funding** This work was supported in part by The National Natural Science Foundation of China NO.61701222.

**References**

1. Bissoto A, Valle E, Avila S (2019) The Six Fronts of the Generative Adversarial Networks, arXiv preprint arXiv:1910.13076
2. Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK (2016) Pap smear image classification using convolutional neural network. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, pp 1–8
3. Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK (2017) Automated classification of pap smear images to detect cervical dysplasia. Comput Methods Prog Biomed 138:31–47
4. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp 213–229
5. Chankong T, Theera-Umpon N, Auephanwiriyakul S (2014) Automatic cervical cell segmentation and classification in pap smears. Comput Methods Prog Biomed 113:539–556
6. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2020) Pre-trained image processing transformer, arXiv preprint arXiv:2012.00364
7. Chouhan N, Khan A, Shah J, Hussnain M, Khan MW (2021) Deep convolutional neural network and emotional learning based breast Cancer detection using digital mammography. Comput Biol Med 132:104318
8. Conceição T, Braga C, Rosado L, Vasconcelos MJM (2019) A review of computational methods for cervical cells segmentation and abnormality classification. Int J Mol Sci 20:5114
9. Denton EL, Chintala S, Fergus R (2015) Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances in neural information processing systems, pp 1486–1494
10. Dong N, Zhai M-d, Zhao L, Wu CH (2020) Cervical cell classification based on the CART feature selection algorithm. J Ambient Intell Humaniz Comput 12:1–13
11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020) An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929
12. Dounias G, Bjørregaard B, Jantzen J, Tsakonas A, Ampazis N, Panagi G, Panourgias E (2006) Automated identification of cancerous smears using various competitive intelligent techniques. Oncol Rep 15:1001–1006
13. Esser P, Rombach R, Ommer B (2020) Taming transformers for high-resolution image synthesis, arXiv preprint arXiv:2012.09841
14. Gautam S, Jith N, Sao AK, Bhavsar A, Natarajan A (2018) Considerations for a PAP smear image analysis system with CNN features, arXiv preprint arXiv:1806.09025
15. Ghoneim A, Muhammad G, Hossain MS (2020) Cervical cancer classification using convolutional neural networks and extreme learning machines. Futur Gener Comput Syst 102:643–649
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
17. Guissous AE (2019) Skin Lesion Classification Using Deep Neural Network, arXiv preprint arXiv: 1911.07817
18. Gv KK, Reddy GM (2019) Automatic classification of whole slide pap smear images using CNN with PCA based feature interpretation. CVPR Workshops, pp 1074–1079
19. Han Y, Wang G (2020) Skeletal bone age prediction based on a deep residual network with spatial transformer. Comput Methods Prog Biomed 197:105754
20. Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa K, Li Y, Nakayama H (2019) Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. In: 2019 International Conference on 3D Vision (3DV). IEEE, pp 729–737
21. Haryanto T, Sitanggang IS, Agmalaro MA, Rulaningtyas R (2020) The Utilization of Padding Scheme on Convolutional Neural Network for Cervical Cell Images Classification. In: 2020 International conference on computer engineering, network, and intelligent multimedia (CENIM). IEEE, pp 34–38
22. He X, Chen Y (2019) Optimized input for CNN-based hyperspectral image classification using spatial transformer network. IEEE Geosci Remote Sens Lett 16:1884–1888
23. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637
24. Hussain E, Mahanta LB, Borah H, Das CR (2020) Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data Brief 30:105589
25. Ibtihaz N, Rahman MS (2020) MultiResUNet: rethinking the U-net architecture for multimodal biomedical image segmentation. Neural Netw 121:74–87
26. Jaiswal A, Gianchandani N, Singh D, Kumar V, Kaur M (2020) Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. J Biomol Struct Dyn 39:1–8
27. Jantzen J, Norup J, Dounias G, Bjervegaard B (2005) Pap-smear benchmark data for pattern classification. In: Nature inspired Smart Information Systems (NiSIS 2005), pp 1–9
28. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA Cancer J Clin 61:69–90
29. Jie H, Li S, Samuel A, Gang S, Enhua W (2019) Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
30. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196
31. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410
32. Khamari A, Gupta D, de Albuquerque VHC, Sangaih AK, Jhaveri RH (2020) Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. J Supercomput 76:1–19
33. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2021) Transformers in Vision: A Survey, arXiv preprint arXiv:2101.01169
34. Kim K, Naylor B (2012) Practical guide to surgical pathology with cytologic correlation: a text and color atlas. Springer Science & Business Media
35. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst 25:1097–1105
36. Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983
37. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101
38. Mamunur Rahaman M, Li C, Yao Y, Kulwa F, Wu X, Li X, Wang Q (2021) DeepCervix: A Deep Learning-based Framework for the Classification of Cervical Cells Using Hybrid Deep Feature Fusion Techniques, arXiv e-prints, arXiv: 2102.12191
39. Marinakis Y, Doungias G, Jantzen J (2009) Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. Comput Biol Med 39:69–78
40. Papanicolau GN, Traut HF (1941) The diagnostic value of vaginal smears in carcinoma of the uterus. Am J Obstet Gynecol 42:193–206
41. Park S, Kim G, Oh Y, Seo JB, Lee SM, Kim JH, Moon S, Lim J-K, Ye JC (2021) Vision transformer for COVID-19 CXR Diagnosis using Chest X-ray Feature Corpus, pp arXiv:2103.07055
42. Peng G, Dong H, Liang T, Li L, Liu J (n.d.) Diagnosis of cervical precancerous lesions based on multimodal feature changes. Comput Biol Med:104209
43. Plissiti ME, Dimitrakopoulos P, Sifkas G, Nikou C, Krikoni O, Charchanti A (2018) SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp 3144–3148

44. Pollastri F, Bolelli F, Paredes R, Grana C (2020) Augmenting data with GANs to segment melanoma skin lesions. Multimed Tools Appl 79:15575–15592

45. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434

46. Ramesh S, Sasikala S, Paramanandham N (2021) Segmentation and classification of brain tumors using modified median noise filter and deep learning approaches. Multimed Tools Appl 80:11789–11813

47. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A (2015) M.J.I.j.o.c.v. Bernstein. Imagenet large scale visual recognition challenge 115:211–252

48. S. McGuire, World cancer report (2014) Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO press, 2015. Adv Nutr 7(2016):418–419

49. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2234–2242

50. Shah V, Autee P, Sonawane P (2020) Detection of Melanoma from Skin Lesion Images using Deep Learning Techniques. In: 2020 International Conference on Data Science and Engineering (ICDSE). IEEE, pp 1–8

51. Shi J, Wang R, Zheng Y, Jiang Z, Zhang H, Yu L (2021) Cervical cell classification with graph convolutional network. Comput Methods Prog Biomed 198:105807

52. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin

53. Talo M (2019) Diagnostic classification of cervical cell images from pap smear slides. Acad Perspect Proc 2:1043–1050

54. Thuy MBH, Hoang VT (2019) Fusing of deep learning, transfer learning and Gan for breast cancer histopathological image classification. In: International Conference on Computer Science, Applied Mathematics and Applications. Springer, pp 255–266

55. Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, pp 242–264

56. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM (2021) Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. arXiv preprint arXiv:2102.10662

57. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762

58. Wang Q, Liu S, Chanussot J, Li X (2018) Scene classification with recurrent attention of VHR remote sensing images. IEEE Trans Geosci Remote Sens 57:1155–1167

59. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA (2020) CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

60. Wang Q, Huang W, Xiong Z, Li X (2020) Looking closer at the scene: multiscale representation learning for remote sensing image scene classification. IEEE Transactions on Neural Networks and Learning Systems

61. Wang L, Zhang C, Bai R, Li J, Duan H (2020) Heck reaction prediction using a transformer model based on a transfer learning strategy. Chem Commun 56:9368–9371

62. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, arXiv preprint arXiv:2102.12122

63. Wieslander H, Forslid G, Bengtsson E, Wahlyc B, Hirsch J-M, Runow Stark C, Kecheril Sadanandan S (2017) Deep convolutional neural networks for detecting cellular changes due to malignancy. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 82–89

64. William W, Ware A, Basaza-Ejiri AH, Obungoloj C (2018) A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Comput Methods Prog Biomed 164:15–22

65. Win KP, Kitjaidure Y, Hamamoto K, Myo Aung T (2020) Computer-assisted screening for cervical Cancer using digital image processing of pap smear images. Appl Sci 10:1800

66. Wu M, Yan C, Liu H, Liu Q, Yin Y (2018) Automatic classification of cervical cancer from cytological images by using convolutional neural network, bioscience reports, 38
69. Xie Y, Zhang J, Shen C, Xia Y (2021) CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation, arXiv preprint arXiv:2103.03024
70. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Tay FE, Feng J, Yan S (2021) Tokens-to-token vit: Training vision transformers from scratch on imagenet, arXiv preprint arXiv:2101.11986
71. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 586–595
72. Zhao L, Li K, Wang M, Yin J, Zhu E, Wu C, Wang S, Zhu C (2016) Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search MRF. Comput Biol Med 71:46–56
73. Zhao C, Shuai R, Ma L, Liu W, Hu D, Wu M (2021) Dermoscopy image classification based on StyleGAN and DenseNet201. IEEE Access 9:8659–8679
74. Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable DETR: Deformable Transformers for End-to-End Object Detection, arXiv preprint arXiv:2010.04159

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chen Zhao Master, College of Computer Science and Technology. His major research topics include image processing and machine learning.
Renjun Shuai  Associate professor, Director of the Institute of Digital City and Intelligent Architecture, Nanjing Tech University. Research directions include intelligent architecture and medical images.

Li Ma  Master, Nanjing Health Information Center. His major research topics include image processing and machine learning.
**Wenjia Liu** Ph.D. associate chief physician, M.D., Deputy Director of Gastroenterology Department, The Affiliated Changzhou No.2 People’s Hospital of Nanjing Medical University.

**Menglin Wu** Ph.D. associate professor, master’s supervisor. Research directions include medical image pathological analysis, medical information retrieval and mining, deep learning application technology.