Down Syndrome Detection with Swin Transformer Architecture

Chengyu Wang[1,2], Limin Yu[3], Jionglong Su[4], Trevor Mahy[5], Valerio Selis[2], Chunxiao Yang[6], and Fei Ma[1] *

Abstract

Objective: Down Syndrome, also known as Trisomy 21, is a severe genetic disease caused by an extra chromosome 21. For the detection of Trisomy 21, despite those statistical methods have been widely used for screening, karyotyping remains the gold standard and the first level of testing for diagnosis. Due to karyotyping being a time-consuming and labour-intensive procedure, Computer Vision methodologies have been explored to automate the karyotyping process for decades. However, few studies have focused on Down Syndrome detection with the Transformer technique. This study develops a *Down-Syndrome-Detector (DSD)* architecture based on the Transformer structure, which includes a segmentation module, an alignment module, a classification module, and a Down Syndrome indicator. *Methods:* The segmentation and classification modules are designed by homogeneous transfer learning at the model level. Transfer learning techniques enable a network to share weights learned from the source domain (e.g., millions of data in ImageNet) and optimize the weights with limited labelled data in the target domain (e.g., less than 6,000 images in BioImLab). The AlignModule is designed to process the segmentation output to fit the classification dataset, and the Down Syndrome Indicator identifies a Down Syndrome case from the classification output. *Results:* Experiments are first performed on two public datasets BioImLab (119 cases) and Advanced Digital Imaging Research (ADIR, 180 cases). Our performance metrics indicate the good ability of segmentation and classification modules of *DSD*. Then, the DS detection performance of *DSD* is evaluated on a private dataset consisting of 1084 cells (including 20 DS cells from 2 singleton cases): 90.0% and 86.1% for cell-level TPR and TNR; 100% and 96.08% for case-level TPR and TNR, respectively. *Conclusion:* This study develops a pipeline based on the modern Transformer architecture for the detection of Down Syndrome from original metaphase micrographs. Both segmentation and classification models developed in this study are assessed using public datasets with commonly used metrics, and both achieved good results. The *DSD* proposed in this study reported satisfactory singleton case-specific DS detection results. *Significance:* As verified by a medical specialist, the developed method may improve Down Syndrome detection efficiency by saving human labor and improving clinical practice.

Keywords: Karyotyping, Down Syndrome, Transformer, Deep Neural Network.

1 Introduction

Down Syndrome (DS) occurs in approximately one in 700 births worldwide. It was first clinically described by Langdon Down in 1866 [1]. The clinical manifestations include intellectual disabilities, cardiac diseases, physical abnormalities, and other abnormalities [2]. First reported as the underlying genomic abnormality in 1959, DS is a genetic disorder in human chromosomes caused by an extra copy of genes on chromosome 21. The content of redundant chromosomal is produced in three different ways [3] : 95% of DS cases are Trisomy 21 (T21), in which all cells have three chromosomes 21 instead of two. 2% of DS cases are called mosaic Down Syndrome, which is diagnosed by the mixture of cells, with some having two chromosomes 21 and some had three. 3% of DS cases are named Translocation Down Syndrome, caused by a partial copy of chromosome 21, which attaches to another chromosome. Since there is no cure for DS so far, besides improving the quality of life through proper care and education, the most efficient way for DS prevention is through screening or diagnostic tests at an early prenatal age [4].

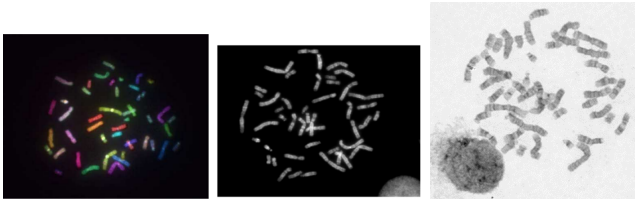
Statistical methods have been used in the noninvasive prediction of chromosomal abnormalities for years. These approaches produced a likelihood percentage of a fetus being suffering a fatal aneuploidy disease [5]. Nicolaides et al. calculated T21 risks using a multivariate likelihood approach in 75821 singleton pregnancies in 2005 [6], reporting a detection rate of 75% with the false-positive rate of 1%. These methods compute the patient-specific risk based on several markers from an antenatal test, such as crown ramp length and fetal nuchal translucency. Neocleous et al. used a fully connected feedforward artificial neural network to predict the risk of T21 and other chromosomal aneuploidies in 2016 [5]. The dataset consists of 51,208 singleton pregnancy cases from first-trimester aneuploidy screening. The screening report was used as suitable markers for aneuploidies risk establishment by nine parameters, including maternal age, previous pregnancy

*Corresponding author: fei.ma@xjtlu.edu.cn
[1]School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China
[2]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, Merseyside L69 3GJ UK
[3]Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
[4]School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, China
[5]School of Languages, Xi'an Jiaotong-Liverpool University, Suzhou, China
[6]Suzhou Precision Medical Technology Co., LTD, Suzhou, China
This work was supported in part by the National Natural Science Foundation of China (Grant No. 61501380), the Qing Lan Project of Jiangsu Province, Laboratory of Computational Physics under Grant 6142A05180501, Xi'an Jiaotong-Liverpool University (XJTLU) Research Development Fund RDF-17-02-51, XJTLU research enhancement fund REF-19-01-04 and REF-18-01-04, Key Programme Special Fund (KSF) in XJTLU KSF-E-32, KSF-E-21, KSF-A-22, and XJTLU Construction of a Bioinformatics Platform for Precision Medicine: RDS10120180041.

with T21, etc. They achieved a 100.0% detection rate of T21 with a false positive rate of 3.9%. Feng et al. devised a nine-layer Convolutional Neural Network (CNN) model consisting of two merged branches for accurate DS prediction/screening in 2017 [2]. Each branch contains three convolutional layers, one max-pooling layer, and one input layer fed with a chromosome single-nucleotide polymorphisms (SNPs) map. The SNPs dataset was built from 378 samples collected by Vanderbilt University Medical Center. Each sample contains the intensity information of 5458 SNPs in 321 HSA21 coding genes. The proposed CNN model achieved an average accuracy of 99.3%, a precision of 99.2%, and a recall of 98.4%, surpassing three conventional supervised learning algorithms (SVM, Random Forest, and Decision Tree) in accuracy (they were all below 97.1%) on the same dataset.

Most statistical approaches calculate either the appropriate multivariate or posterior probabilities from markers such as fetus crown ramp length, nuchal translucency, maternal age, the pregnancy-associated plasma protein-A, etc [5]. Karyotyping, however, is generally recognized as the gold standard in diagnosing genetic abnormalities in fetuses via checking chromosomal abnormalities of numerical (an extra or missing chromosome) or structural (deletion, translocation, inversion in specific chromosome segments) [2, 7]

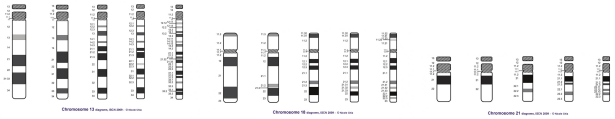
Each human chromosome contains a single deoxyribonucleic acid (DNA) duplex [8] that is only visible during mitosis or meiosis via techniques such as M-FISH (multicolour fluorescence in situ hybridization with five colour dyes [9]), Q-Band (staining with the fluorescent dye nitrogen mustard quinacrine [10]), or G-Band (stained with the dye of Giemsa [11]). Examples of the chromosome images, obtained from the three techniques, showing the different Band styles, are given in Figure 1.



(a) M-FISH (b) Q-Band (c) G-Band

Figure 1: Examples of chromosomal images obtained by different techniques. (a), (b), and (c) are from the dataset of ADIR, BioImLab, and a private dataset, respectively.

Karyotyping is the process of preparing karyotypes from photographs of chromosomes to detect the numerical and structural abnormalities of the cell. It is widely used for specific cancer diagnoses and prenatal screening of several genetic diseases [12]. Since the 1980s, karyotyping has been carried out for prenatal screening in the first trimester of gestation by chorionic villus sampling [13]. After the identification of chromosomes in a photomicrograph (an example is shown in Figure 1), each chromosome is then compared to the idiogram and is assigned with a label from 1 to 24. The idiogram is the phenotypic representation of the chromosomal centromere and bands. Figure 2 gives an example of an idiogram. It corresponds to the karyotype images obtained by typical staining techniques of Q-banding, G-banding and R-banding (the reverse of G-banding) [14]. Initially published as part of the International System for Human Cytogenetic Nomenclature (ISCN) in 1971 and most recently revised in 2020 [15], the idiograms demonstrate relative centromere position and banding patterns for 24 chromosome types by a series of bands in white and black [16]. Karyotyping is a time-consuming and labor-intensive task highly dependent on skilled clinical analysts; therefore, automatic methods have been researched to reduce the burden [17] [18].



(a) chromosome 13 (b) chromosome 18 (c) chromosome 21

Figure 2: Examples of idiogram.

Computer-aided karyotyping methods involve four tasks: pre-processing, segmenting a chromosomal micrograph into individual chromosomes, classifying each chromosome using one of the 23 class labels, and abnormal detection [19]. Traditional segmenting methods are mainly based on cut points that extract morphological characters [7], e.g., OTSU, thresholding, K-means clustering, watershed. The typical workflow consists of pre-processing the images, detecting contours, drawing cut lines, and segmenting the potential homologous chromosomes. Neural Network-based automatic chromosome karyotyping can be traced back to the 1990s. Boaz Lerner investigated a multilayer perceptron Neural Network for chromosome analysis in 1998 [20], achieving an accuracy of 83.6%, the best performance at that time.

With the breakthrough of Deep Learning in 2012, characterized by using the deep convolutional network [21], the efficiency and accuracy of computer-aided chromosome karyotyping was dramatically improved. Transformer architecture, which has already been dominant in Natural Language Processing for years, has ushered Computer Vision into a new phase in 2020 [22–24]. Several researchers evaluated their methods of chromosome classification or segmentation on the public chromosome image datasets, e.g., the Laboratory of Biomedical Imaging from the University of Padova (BioImLab, Q-Band technique [10, 25]) and Advanced Digital Imaging Research from the University of Texas at Austin (ADIR, M-FISH technique [9, 26]):

- **BioImLab**: (1) Grisan *et al.* proposed a space-variant thresholding scheme to separate chromosomes in 2009 [10], they obtained 94% correctly segmented chromosomes; (2) Grisman *et al.* improved their research by a region-based level set algorithm to deal with the image background [27], with 98.0% and 81.0% identification accuracy for a single and over-segmented chromosomes, respectively; (3) Poletti *et al.* implemented a thorough analysis of the eleven thresholding methods, which achieved better performance than that in 2012 [28].
- **ADIR (Advanced Digital Imaging Research)**: (1) Schwartzkopf *et al.* proposed a hypothesis test strategy based on maximum-likelihood in 2005 [9], achieving an accuracy of 77% on segmenting touching chromosomes and 34% on overlapping chromosomes. (2) Karvelis *et al.* investigated the watershed transform and gradient paths in 2010 [29]. The proposed

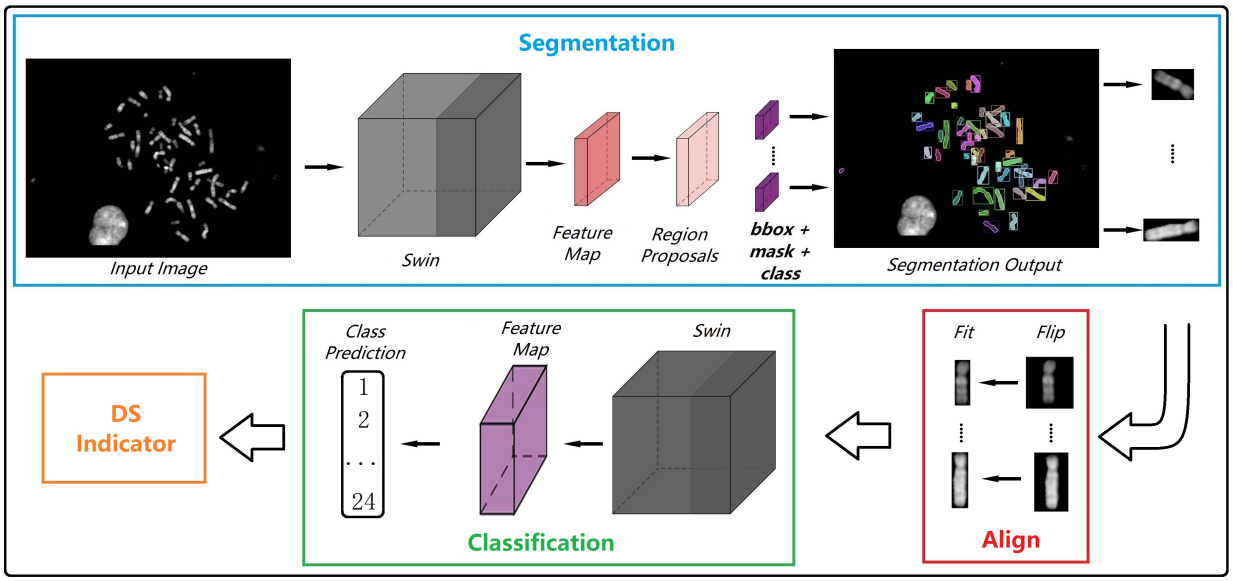


Figure 3: The overall architecture of Down-Syndrom-Detector.

algorithm achieved a success rate of 90.6% and 80.4% for touching and overlapping chromosomes, respectively. (3) Pardo *et al.* studied a fully convolutional network with a VGG-alike layout for semantic segmentation in 2018 [26], yields 87.41% for average correct classification ratio. (4) Arora and Tanvi *et al.* introduced a novel strategy in 2019 [30], which employs a Gaussian kernel (energy function) to fit the foreground and background regions. They validated their work on 66 images from three m-FISH (ADIR), Q-Band, and G-Band datasets; each of the datasets contains 22 images. The accuracy of corrected segmented chromosomes is **96.33%**.

Xiao et al. studied object detection schemes and developed a DeepACEv2 for chromosome enumeration task in 2020 [31]. The framework adds Hard Negative Anchors Sampling to extract information from confusing partial chromosomes and extract unique embeddings proposals from geometric chromosome information by a Template Module branch of each proposal. Experiments were implemented on 1375 metaphase images and reported the Whole Correct Ratio of 71.39% for images. However, their work was not validated on public datasets such as BioImLab (with only 163 metaphase images). There was also no reporting on the results of specific Trisomy such as T21.

Al-Kharraz et al. proposed an automated karyotyping workflow with Deep Learning in 2020 [32] to recognize numerical abnormalities of Trisomy 13/21/18/sex chromosome. Their experimental dataset contains 147 non-overlapped metaphase chromosome images from the Center of Excellence in Genomic Medicine Research at King Abdulaziz University. The network YOLOv2 [33] is used to classify the individual chromosomes based on network VGG19 [34]. Twenty-nine metaphase cells were used for abnormal detection. They reported 96.6% abnormality detection accuracy, defined as the ratio of the correctly diagnosed cell number divided by the total cell number. Al-Kharraz et al. reported the classification results on the segmentation output of the dataset (CEGMR) with only non-overlapped images. However, the results on more challenging datasets such as BioImLab are not given.

Although there have been many successful Transformer architectures in Computer Vision since 2020 [23] [35] [36], they have not been used for karyotyping. To the best of our knowledge, this paper is the first study to adopt the Transformer architecture for Down Syndrome detection. The rest of this paper is organized as follows: Section 2 reviews Transformer techniques in Computer Vision; Section 3 explains the proposed DS detection workflow and the transfer learning strategies; Section 4 introduces the experiment design, datasets, and metrics; Section 5 gives the results and discussion. Finally, the conclusion is given in Section 6.

2 Transformer Networks for Computer Vision

Cho and Bengio first proposed the attention mechanism for the neural machine translation task in 2016 by focusing only on the relevant source-target word [37]. The conditional probability in the decoder is defined as: $p(y_i|y_1,...,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i)$ where $s_i = f(s_{i-1},y_{i-1},c_i)$ is a hidden state for time i in Recurrent Neural Net (RNN), in which it only calculates a context vector $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$ for each y_i, instead of the whole vector c in RNN. Vaswani *et al.* devised a new building block known as (the standard) Transformer in 2017 [22], which solely uses self-attention layers instead of convolutions layers and RNN architecture. The main ingredients in the encoder and decoder are the stacked six identical blocks. Given the queries Q, keys K, and values V with dimension d_k, the attention output is calculated as [22]

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In the field of image classification, Dosovitskiy *et al.* applied the position embedded image patches to the Transformer [22], and devised a Vision Transformer in 2017 [23]. To supplement the missing information of context, 2D interpolation was performed during position embeddings via the location of the original images.

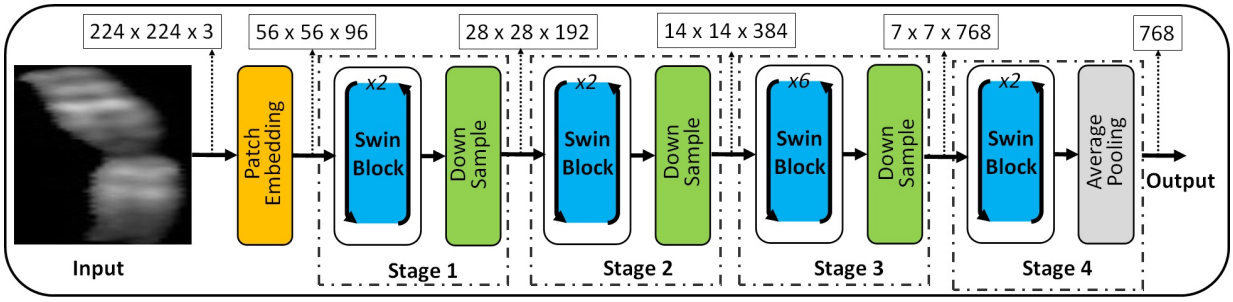


Figure 4: The structure of the Swin Transformer tiny version used in the *Classification Module of DSD*: The height and width are halved in the *Down Sample* operation, while the channel is doubled. The only difference in the Swin Transformer structure of the *Segmentation Module of DSD* is that there are 18 blocks in the third stage instead of 6. The number $h \times w \times c$ of the top indicates the output dimension in the corresponding step.

Carion *et al.* proposed the DETR method with Transformers for object detection tasks [24], using the parallel Transformer [23] and bipartite matching to output the final predictions set in parallel. A traditional CNN backbone learns the 2D representations (e.g., ResNet 50 [38]), then flattens with positional embeddings, and is finally inputted into a Transformer encoder to output the prediction as an object or 'no object' with a shared Feed Forward Network. DETR does not require post-processing such as Non-Maximum Suppression, which is obligatory for modern detectors (e.g., Faster R-CNN [39]), and reports comparable results to the baseline Faster R-CNN, especially on large objects. However, DETR requires much more training time than traditional CNN architecture to converge [24].

Zhu *et al.* improved DETR as Deformable DETR to address the issue of slow convergence and limited resolution for feature space by using a deformable attention module to replace the attention module in DETR [40]. Four-scale feature maps are used as input to aggregate multiscale feature maps in the encoder. An embedding is added for each scale level. This process is similarly implemented via Feature Pyramid Network alike structure, which benefits modern detectors. In the decoder, the (multi-scale) deformable attention is the core of Deformable DETR, which focuses on a small set around a reference point.

Sun *et al.* developed a Sparse R-CNN by learning a fixed sparse (e.g., 100) proposal set to refine the dense object candidates (up to hundreds of thousands bounding box proposals) [41]. Given an input image, the region proposal boxes and proposal features are learned from the dynamic head to output the predicted object location and class label. Each proposal box corresponds to a proposal feature extracted individually by implementing the RoI Align operation in the dynamic head. A bipartite matching loss optimizes the predictions with a ground truth label. The proposed Sparse R-CNN illustrates performance on par with the well-established detector baselines on the COCO dataset.

Liu *et al.* proposed a Swin Transformer to bridge the gap of a unified backbone in both vision and language tasks [36]. A hierarchical Transformer architecture on the basis of the scheme of shifted windows is proposed, limiting the computation within a small patch instead of the whole feature map. The patches are split from hierarchical scale features and gradually merge between neighbours in deeper layers. By replacing the backbone (e.g., ResNet 50) in typical object detection/segmentation frameworks (e.g., Mask R-CNN) with the proposed Swin Transformer blocks, Liu *et al.* reported state-of-the-art performance of image detection/segmentation task on ImageNet.

Previous studies on Transformer have focused on generic datasets (e.g., ImageNet, COCO). However, the differences in image styles and feature distributions between generic and medical-specific datasets (e.g., chromosome images) are valuable and can be further explored. This study explores them by predicting Down syndrome with a transfer learning strategy.

3 Methods

An overview of the proposed *Down-Syndrome-Detector (DSD)* workflow is given in Figure 3. From left to right: (1) The first (blue) component is the *Segmentation Module of DSD* which recognizes each potential chromosome instance from a whole microscopical image and draws a bounding box and masks for each chromosome. The instance number of ground truth is 46 for a normal cell and 47 for a DS cell. (2) The second (red) part is the *Align Module of DSD* which firstly rotates a chromosome image vertical, then removes the area outside the chromosome border. (3) The third (green) constituent is the *Classification Module of DSD* which identifies the belonging category of the input chromosome image, which is one of 24. (4) The last (orange) part is the *Down Syndrome Indicator module of DSD* which detects a DS case by counting the number of chromosome 21 via:

$$I = \begin{cases} 1, \text{ if } \dfrac{\mathrm{N} - \mathrm{N_T}}{\mathrm{N_T}} < \mathrm{F}, \text{ when } \mathrm{N_T} > 0 \\ 0, \text{ if } \dfrac{\mathrm{N} - \mathrm{N_T}}{\mathrm{N_T}} \geqslant \mathrm{F}, \text{ or } \mathrm{N_T} = 0 \end{cases} \tag{2}$$

where I is the DS indicator for a singleton case consisting of N cells. A cell with more than two chromosomes 21 is identified as a T21 cell. N_T is the number of T21 cells. F is the *DS Indicator Factor*. In general, a singleton case contains around ten cells. In particular case, only one cell is available.

In our proposed *DSD*, Swin Transformer is used as the backbone module of the segmentation and the classification module [36]. The attention mechanism in Swin Transformer enables the network to add more weights to the context of interest, achieving higher performance and lower parameter cost. Figure 4 gives the structure of the Swin Transformer. First, an input image of size $224 \times 224 \times 3$ is partitioned into patch tokens of resolution

$h0 \times w0 \times c0$ with patch size 4×4 and $c0 = 96$, where $h0 = \frac{224}{4}, w0 = \frac{224}{4}$. Both $c0$ and patch size 4×4 are recommanded values for the tiny version of Swin Transformer [36]. The patches then go through four similar stages in turn, with output channels of 192, 384, 768, and 768. Finally, a Fully Connected layer maps the feature map to a 24-dimensional class prediction.

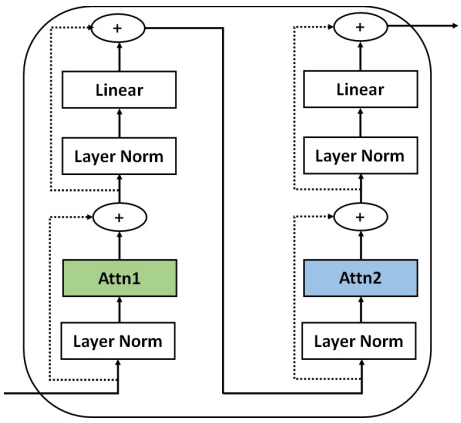


Figure 5: The structure of *Swin block*. The activation after the *Linear* layer is GELU.

Figure 5 gives the structure of the *Swin block* used in the stage 1 to stage 4 in Figure 4. As a comparison, Multihead Self-Attention (MSA) [22] calculates the relationship between a patch and all other patches, whereas *Attn1* performs Window-based Multihead Self-Attention (W-MSA) [36] only between local windows for faster performance. *Attn2* is an upgrade operation of *Attn1*, which strengthens the connection between these non-overlapping windows by shifting the windows down one pixel to the right. Assuming that an image is partitioned into $h \times w$ non-overlapping patch tokens and each window contains $M \times M$ tokens, MSA and W-MSA are computed as follows [36]:

$$\begin{aligned} \Omega(\text{MSA}) &= 4hwC^2 + 2(hw)^2C \\ \Omega(\text{W-MSA}) &= 4hwC^2 + 2M^2hwC \end{aligned} \tag{3}$$

Humans can transfer knowledge from one domain to another. For example, it is much easier for someone to ride a motorbike if he or she has previously learned to cycle. In Deep Learning, homogeneous transfer learning refers to the techniques of improving the model performance on a limited dataset (target domain) by using a learned model on a source domain that consists of sufficient data, to make accurate classification results on the target domain [42] [43]. Previous studies categorize transfer learning into rational-based, data-based (instance-based and feature-based), and model-based (parameter-based) approaches.

Parameter sharing is an intuitive strategy of the model-based approaches which focuses on sharing the parameters between a source domain and a target domain. In this work, the hierarchical Vision Transformer model based on the blocks of Figure 5, i.e., Swin Transformer [36], is used as a backbone to extract features from an input image, then the features are used for instance segmentation or chromosome classification, as shown in Figure 3. The initial parameters are the weights which pre-trained on the source domain (ImageNet [44]), and then the feature representations are further learned from the target domain (BioImLab) as follows:

$$A_t = A_s + A_c \tag{4}$$

where A are the attention parameters in equation 1. The subscripts t, s, and c denote the parameters from the target domain, the source domain, and the difference between them, respectively. The A_c also contains parameters from the fully connected layer with an output dimension of 24, which is the number of chromosome categories. It is worth mentioning that in addition to the Swin Transformer backbone, *Segmentation Module of DSD* includes two other components: the region proposal network for predicting bounding boxes and the RoI alignment module for predicting masks. The parameter optimization methods of these two components are similar to those of the backbone.

4 Experiments

This research investigates the performance of:

- *Segmentation Module of DSD* on two public datasets of BioImLab and ADIR. In addition to Swin Transformer, three popular CNN architectures with segmentation/detection baselines (Faster R-CNN, Mask R-CNN, and RetinaNet) are evaluated for comparison with two modern transformer models (Deformable DETR, Sparse R-CNN). The results are given in Section 5.1.

- *Classification Module of DSD* on BioImlab and a private dataset. First, Section 5.2 gives the 24-class classification performance. Then, Section 5.3.1 explores the DS detection capability. The classification baselines of two popular CNN architectures (ResNet 50 and SE-ResNeXt 50 [45]) are evaluated for comparison.

- *Down-Syndrome-Detector* workflow on a private dataset with 20 real DS cells from two singletons. The results are given in Section 5.3.2

Experiments are conducted under a server that equips four GPUs of NVIDIA GeForce RTX 2080 Ti and is installed with an Ubuntu OS with version 20.04.3 LTS.

4.1 Datasets and Trisomy 21 data

An original metaphase microphotograph is used for segmentation, and then the chromosomes isolated from it are used for classification. A karyotype image includes all the chromosomes that have been sorted and arranged orderly. For a normal cell, there are 46 chromosomes. This study tags image samples from segmentation and classification datasets with an identical cell ID if they come from the same human cell.

Three datasets are used in this study, and examples are given in Figure 1: (1) BioImLab is a publicly available Q-Band chromosome dataset containing images from 163 cells for segmentation (metaphase images) and classification (karyotype and individual chromosome images). (2) ADIR is also publicly available and shares 200 M-FISH metaphase images, in which 17 images are reported as 'difficult to karyotype'. (3) The private dataset is provided by a company in Suzhou, China, and contains images from

1084 cells. Of these, 243 metaphase images are used to evaluate the DS detection performance of the *DSD* workflow, and the remaining images from 841 cells are used for segmentation and classification training and testing.

Since both public datasets do not contain DS cases and the private dataset contains only two, 25% of the cells are randomly selected from the classification dataset, and a chromosome 21 is copied to constitute an additional chromosome 21. For obtaining the T21 cells in the segmentation test set, one chromosome 21 is randomly selected from the karyotype image corresponding to the T21 cells in the classification test set and pasted into the corresponding microscope image. Both the segmentation and the classification dataset are empirically split into three subsets, i.e., the training set (85%), the validation set (5%), and the test set (10%). The numbers of samples in each subset are listed as follows:

Table 1: Number of samples in the sub-sets.

sub-set	ADIR	BioImLab	Private
training	153	95	714
validating	9	8	42
testing	18	16 (4)	85 (21)
for *DSD*	0	0	**243 (20)**
total	180	119	1084

The numbers in parentheses are the number of T21 cells, and each T21 cell contains 47 chromosomes. The T21 cells in the BioImLab test set are No. 18, 44, 65, and 89.

Note that 243 cells from 53 singleton cases in the private dataset are used to evaluate the DS detection performance of *DSD*, of which 20 real DS cells are from two singleton cases.

4.2 Metrics

The metrics of Precision, True Positive Rate (TPR, Recall), True Negative Rate (TNR), and Intersection over Union (IoU) are defined as:

$$Precision = TP/(TP + FP)$$
$$TPR = TP/(TP + FN)$$
$$TNR = TN/(TN + FP)$$
$$IoU = \frac{overlapping\ pixel's\ number}{union\ pixel's\ number}$$

where TP, FP, TN, and FN represent True Positive, False Positive, True Negative, and False Negative, respectively. False Positive Rate (FPR) = 1 - TNR.

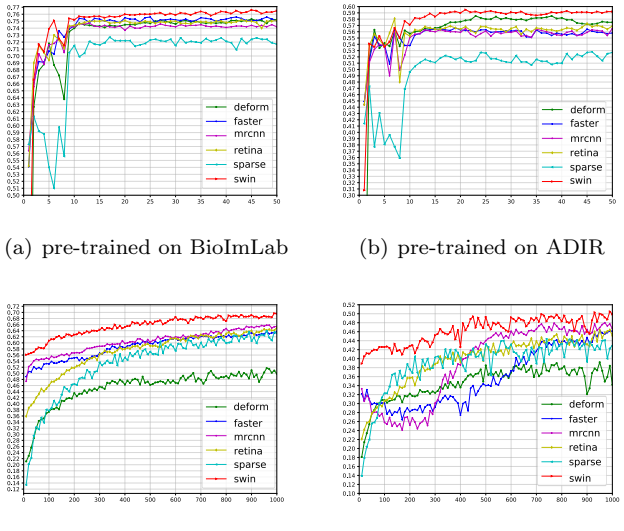
Given a P-R curve which demonstrates the relationship between Precision (P) and Recall (R), mean Average Precision (mAP) is calculated as the area under the P-R curve. The metric mAP is widely used to evaluate the model performance in object detection/segmentation tasks [46] [47]. AP50 and AP75 denote mAP at IoU level of $\geqslant 0.5$ and $\geqslant 0.75$. APs and APm represent $object\ pixels < 32^2$ and $32^2 < object\ pixels < 96^2$, respectively [47]. AP50, AP75, APs, and APm reflect segmentation quality more comprehensively and are widely used to evaluate the Deep Learning models' performance [48]. In medical imaging, the TNR index is important because it relates to the chance of misclassifying an abnormal case as a normal case, which can lead to missed treatment.

5 Results and Discussion

5.1 Segmentation

The results of all models trained with both transfer learning strategy and scratch are given in Table 2, from which we found: (1) All models significantly perform better with a transfer learning strategy (in upper lines) than without (in lower lines). (2) Among 20 evaluation indicators, including with/without transfer learning, two datasets, and five metrics, Swin Transformer outperformed the other models on ten indicators. Its performance is worse than the best results by no more than 1.0% on the other seven indicators. The results indicate the stability and good performance of the Swin Transformer model.

The validation curves of the training process are shown in Figure 6. It is found in Figure 6: (1) The curve of the BioImLab dataset is smoother than that of the ADIR dataset. The results for each model on the BioImLab dataset are at least 10% higher than those on the ADIR, whether pre-trained or trained from scratch. These indicate that the BioImLab data are easier to train. (2) The red curves (Swin Transformer) in (a), (b), (c), and (d) are higher than the other five curves, indicating that Swin Transformer has the best performance and is the most stable. (3) Comparing the segmentation performance with and without transfer learning, the models with transfer learning (a and b) outperform the models without transfer learning (c and d) by more than 5%.



(a) pre-trained on BioImLab (b) pre-trained on ADIR

(c) scratch trained on BioImLab (d) scratch trained on ADIR

Figure 6: Training curves of segmentation models on two datasets: the model names are represented in the same way as in Table 2. The horizontal axis is the number of training epochs, and the vertical axis is the mAP result of the model on that epoch. 'pre-trained' denotes trained with transfer learning technique.

5.2 Classification on 24 Categories

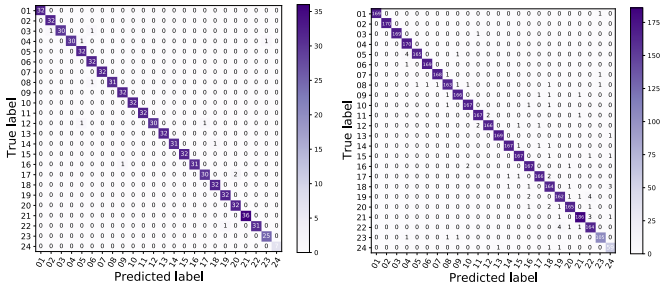
The confusion matrix in Figure 7 depicts the overall performance of the *Classification Module of DSD* in identifying the 24 categories. It is found that the module can correctly classify most chromosomes, suggesting the possibility of diagnosing other trisomy syndromes, such as T18 and T13. Specifically, the accuracy, precision, and

Table 2: Detection performance on BioImLab and ADIR with six models.

	ADIR (%)					BioImLab (%)					Param	Time
model	mAP	AP50	AP75	APs	APm	mAP	AP50	AP75	APs	APm	(M)	(sec.)
faster	68.5	96.3	81.6	69.3	69.0	79.8	98.6	94.5	80.7	74.6	41.12	18
	55.5	87.6	64.2	58.6	40.1	66.4	92.4	79.3	68.9	32.6		
mrcnn	68.8	96.3	83.3	69.3	69.9	79.6	97.8	95.7	80.6	72.3	43.75	21
	58.5	*89.3*	*67.7*	*61.5*	43.1	*69.3*	*94.2*	*80.9*	*71.6*	31.4		
retina	67.9	94.2	78.3	68.8	66.4	79.1	95.5	91.6	80.5	68.5	36.1	19
	55.4	88.1	60.8	58.7	32.0	66.7	91.0	77.5	70.3	30.2		
defor	69.5	96.4	84.2	70.3	67.2	80.6	98.0	94.3	81.6	69.9	40.8	33
	50.5	74.5	57.3	54.3	39.9	52.0	72.4	65.6	56.2	24.5		
spar	64.3	92.2	73.5	67.0	56.4	78.3	95.6	91.4	80.1	63.7	106.0	23
	55.8	82.6	65.2	58.5	*46.5*	65.2	84.8	77.6	70.0	**50.3**		
swin	69.2	95.5	80.1	69.5	69.5	79.6	97.6	94.1	80.6	75.0	68.69	22
	(-0.3)	(-0.9)	(-4.1)	(-0.8)	(-0.4)	(-1.0)	(-1.0)	(-1.6)	(-1.0)	(+0.4)		
	59.8	**91.4**	**68.0**	**61.7**	**52.8**	**71.6**	**95.9**	**87.0**	**73.3**	*45.3*		
	(+1.3)	(+2.1)	(+0.3)	(+0.2)	(+6.3)	(+2.3)	(+1.7)	(+6.1)	(+1.7)	(-5.0)		

Model name faster, mrcnn, retina, defor, spar, and swin denotes Faster R-CNN, Mask R-CNN, RetinaNet, Deformable DETR (the version with iterative bounding box, refinement, and two-stage mechanism), Spars R-CNN (the version *300 proposals*), and Swin Transformer (the version *Small*), respectively. *The results in the upper lines for each model are obtained by transfer learning strategy. The results in lower lines are obtained by models trained from scratch.* The underlined numbers are the best results in the upper lines. The **numbers in bold** and *numbers in italic* are the best and second best results in lower lines. The numbers inside the brackets are the deviation between the best and Swin Transformer results; the red number indicates the deviation is a minus value.
Param and Time denote the number of parameters and inference time, respectively. The inference time is tested with 16 images from the BioImLab dataset, each image in a size of 768 × 576 pixels.

recall for classifying 24 types were 95.27%, 94.86%, and 98.38%, respectively, on BioImLab; and 95.90%, 95.83%, and 97.81%, respectively, on the private dataset.



(a) BioImLab (b) Private dataset

Figure 7: Confusion matrices for 24 categories by the *Classification Module of DSD*. The total number of each autosome (chromosomes 1 to 22) is 32 and 170 in BioImLab and private datasets, respectively. Except for chromosome 21, which is 36 and 191. The number of Y chromosomes is 7 (7 males) and 62 (62 males) in BioImLab and private datasets, and the number of X chromosomes is 25 (9 females and 7 males) and 108 (23 females and 62 males).

5.3 Down Syndrome Detection

5.3.1 Capability of the *Classification Module of DSD*

To evaluate the DS recognition capability of the *Classification Module of DSD*, Swin Transformer is compared with two popular baseline models: ResNet 50 (Res) and SE-ResNeXt 50 (SE), using a 50-layer version. The results of the three models on the classification test set are shown in Table 3: (1) The columns of *'No. 21'* are the result of predicting chromosome 21 out of 740 samples. That is: the true positive samples are only those that are correctly predicted as chromosome 21, ignoring the predictions of other chromosome categories and classifying them as true negative. (2) The columns for *'Down Syndrome'* are predicted from T21 cells, each containing more than two chromosomes 21. It is found in Table 3:

- On BioImLab (upper rows of each model), *Classification Module of DSD* with Swin Transformer (*Swin*) correctly predicted all chromosome 21 and DS cases (bold numbers in table 3), the other two models only reported full scores for TPR.
- On the private dataset (bottom rows of each model), most of the metrics are worse than on BioImLab for all models, except for the five metrics of ResNet 50 (red numbers in table 3). It indicates that Q-band images are easier to be recognized for all three classification models.

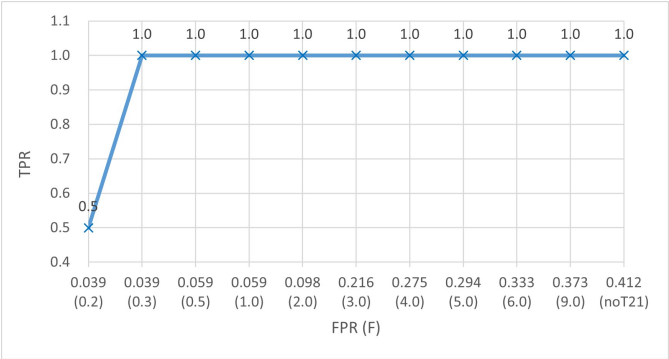


Figure 8: The Sensitive Curve of different *DS Indicator Factor* Values (F). The TPR is mostly high because the test set is composed of unbalanced samples, containing only two positive cases out of 53.

Table 3: Results of DS detection on the classification test set.

model	on Classification Test set (%)							
	NO. 21				Down Syndrome			
	Acc	Preci	TPR	TNR	Acc	Preci	TPR	TNR
Res	99.86	97.30	100	99.86	93.75	80.00	100	91.67
	99.77	97.89	97.38	99.89	95.29	86.96	95.24	95.31
SE	99.86	97.30	100	99.86	93.75	80.00	100	91.67
	99.67	95.41	97.91	99.76	91.76	76.92	95.24	90.62
Swin	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	99.80	98.41	97.38	99.92	96.47	90.91	95.24	96.88

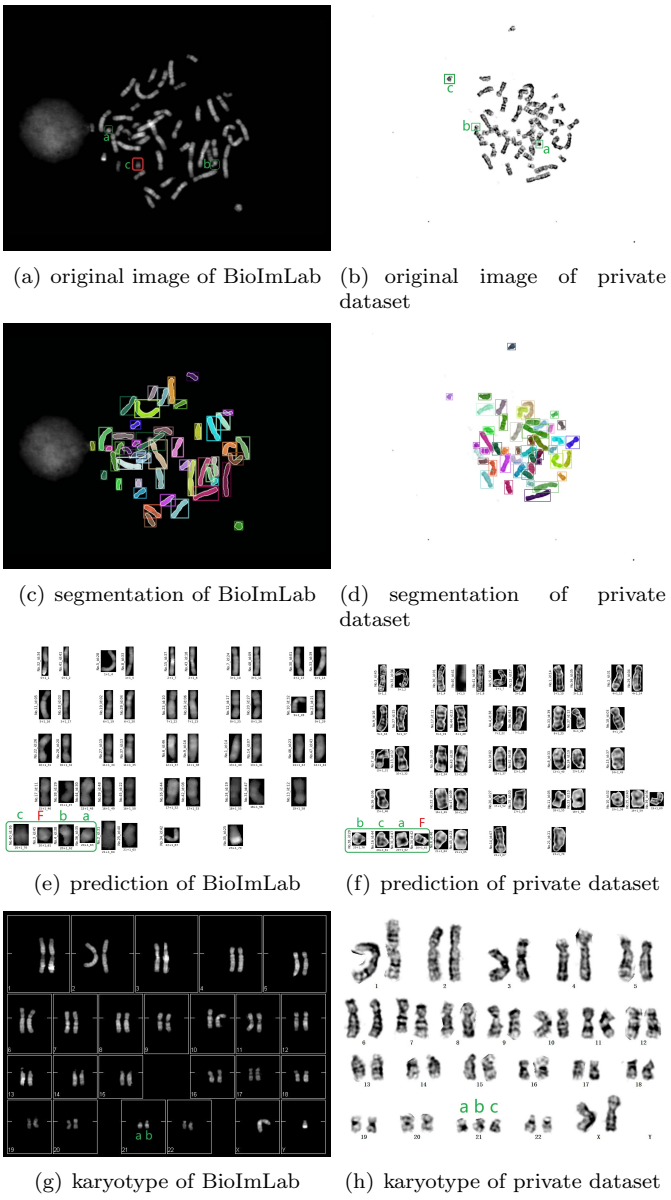
The results of each model's top and bottom row are from BioImLab and private datasets, respectively. For the private test set, the number of T21 cases and the number of total cases is 21 and 85, respectively, which is approximately 5.3 times that of BioImLab (4 T21 cases out of 16 test set cases).

5.3.2 Capability of the Combined *DSD* Workflow

The inputs of the *DSD* workflow are 243 metaphase micrographs from 53 singleton cases. The images are processed according to the process shown in Fig. 3. To pick the appropriate F value in Equation 2 for higher performance, the sensitivity curve in Figure 8 depicts the relationship between TRP and TNR for different values of F. It is found that the highest performance of TPR (1.0) and FPR (0.39) was achieved when $F = 0.3$.

Figure 9 gives examples of the *DSD* workflow: original metaphase micrograph (a is a manufactured T21 cell from BioImLab, b is a true T21 cell from the private dataset), the segmentation output (c and d), the predicted karyotype image by *DSD* (e and f), and the ground truth karyotype image (g and h). The predicted chromosome 21 is placed in the green box at the bottom left of the predicted karyotype image (e and f). It is found that chromosome 21 is smaller than most other chromosomes, and *DSD* prefers to identify these small chromosomes as chromosome 21. While this leads to a higher prediction of false positives, it reduces the prediction of false negatives. False negatives mean that abnormal cases are incorrectly diagnosed as normal and are likely to miss treatment opportunities.

Table 4 gives the DS detection performance of the segmentation output for the segmentation test set on the private dataset. The k-fold validation is implemented with



(a) original image of BioImLab (b) original image of private dataset

(c) segmentation of BioImLab (d) segmentation of private dataset

(e) prediction of BioImLab (f) prediction of private dataset

(g) karyotype of BioImLab (h) karyotype of private dataset

Figure 9: Examples from cell 65 of BioImLab and cell 16 of singleton case 49 of the private dataset. (a) and (b) are the original micrographs. Chromosome 21 is labeled in green and red boxes. The chromosome in the red box in (a) is pasted from the karyotype image of the same cell. (c) and (d) are the outputs predicted by *Segmentation Module of DSD*. (e) and (f) are karyotype images predicted by *Classification Module of DSD* from (c) and (d). (g) and (h) are the corresponding ground truth karyotype images. Letters a, b, c, and F are used to distinguish chromosome 21, and F is an incorrect prediction.

$k = 5$ as follows: (1) All the 841 cases are separated into ten folds, each containing 84 cases except the last one containing 85 cases. (2) Five folds are randomly selected as a testing set, respectively, with the rest consists the training set. (3) The mean and standard deviation values are calculated. It can be found from Table 4 that:

- The *DSD* with the Swin Transformer (*Swin*) as the backbone performs best. It outperforms the other two models in most metrics except the TPR of 243 cells (red numbers). TPR (Recall) and TNR all surpassed 96% (underlined numbers) for DS detection of singleton cases. In particular, the TPR reached 100%.

- For all models, the accuracy, TPR, and TNR are higher for singleton-level than for cell-level. This is more favourable in clinical practice, where DS is diagnosed by singleton cases rather than cells.

- In both table 3 and 4, the Swin Transformer performs better than the other two CNN-based strong baseline models (ResNet 50 and SE-ResNeXt 50). It may be attributed to the attention mechanism of the Transformer structure.

- In the results of **DS Detection from Segment output**, the low precision of all models is partly due to the fact that there are only 20 DS cells out of 243 cells (2 out of 53 singleton cases). Further discussion will be taken with the table 5 in the next paragraph.

Table 5 gives the DS prediction results for the cells included in the DS cases detected by *DSD* with *Swin* in Table 4. Two of the four singleton cases are correctly predicted (bold), and the other two are incorrect (red). It is found that correctly detecting DS singleton cases is challenging if the number of cells contained in the singleton is too small (no more than three).

Table 5: Prediction of DS singleton cases.

Singleton ID	Total	T21	Normal	Factor
3	1	1	0	0
16	3	3	0	0
49	10	8	2	0.25
53	10	10	0	0

The numbers in the middle three columns are the number of cells. The *Factor* is calculated as $\frac{N-N_T}{N_T}$, where N and N_T are the total number of cells and the number of T21 cells, respectively, in the singleton case.

To illustrate which region contributes more to the identification of chromosome 21, Figure 10 and Figure 11 visualize the features in different colors by using Grad Cam++ [49]. Compared to ResNet 50 and SE-ResNeXt 50, Swin Transformer focuses more on the dark regions of chromosomes, which are the banded regions of chromosomes and are crucial for classifying classes of chromosomes.

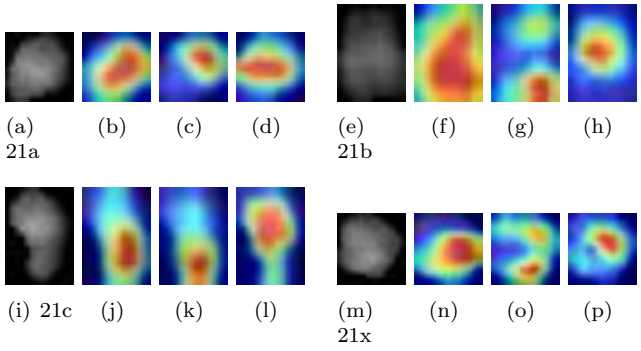


Figure 10: Feature visualization of chromosome 21 of BioImLab case 65. From left to right, four images in a group are the input chromosome image (a,e,i,m), feature visualization from ResNet 50 (b,f,j,n), SE-ResNeXt 50 (c,g,k,o), and Swin Transformer (d,h,l,p). 21a, 21b, and 21c are the correctly predicted chromosome 21. 21x is the misclassified one.

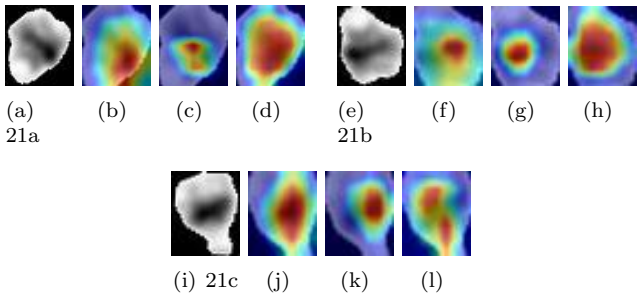


Figure 11: Feature visualization of chromosome 21 of the private dataset case 824. As in Figure 10, the images are also from input (a,e,c), ResNet 50 (b,f,j), SE-ResNeXt 50 (c,g,k), and Swin Transformer (d,h,l) of each group. 21a, 21b, and 21c are all correctly predicted as chromosome 21.

6 Conclusions

This work is the first to propose an integrated workflow (*DSD*) via the transfer learning strategy and modern Transformer technique to predict the DS cases from the original metaphase micrographs. Experiments to assess the ability of *Segmentation Module of DSD* and *Classification Module of DSD* are performed on two public datasets and evaluated by several commonly used metrics, which makes it more feasible to compare this work with future studies. On a private dataset containing 20 real DS cells from two singleton cases, *DSD* reported satisfactory TPR and TNR for detecting DS. Although the True Negative Rate is high (96.08%), the precision of positive samples is low (50%), which may be due to (1) the weak ability to distinguish those chromosomes that are similar in size to chromosome 21. (2) Lack of a sufficient number of cells in singletons cases. These issues should be further investigated in future work. A medical specialist in clinical cytogenetics verifies that the study's findings may improve Down Syndrome detection efficiency by saving human labour and enhancing clinical practice. The medical specialist claims a semi-automatic process of Down Syndrome Detection with a singleton case with ten cells costs more than 400 seconds. Here, a semi-automatic process means the process by a specialist with a software system. An experiment based on the proposed architecture is designed, which runs on a laptop with a GPU 1660ti. In the

Table 4: Results of T21 detection on classification set and DS detection on segmentation output.

model	**T21 Detection on classification set (%)**				**DS Detection from Segment output (%)**							
					of 243 cells				of 53 singleton cases			
	Acc	P	TPR	TNR	Acc	P	TPR	TNR	Acc	P	TPR	TNR
Res	0.9955	0.9422	0.9571	0.9973	0.7333	0.2328	0.8800	0.7202	0.9283	0.4254	**1.0000**	0.9255
	0.0007	*0.0189*	*0.0248*	*0.0010*	*0.0766*	*0.0516*	*0.0509*	*0.0866*	*0.0452*	*0.1983*	*0.0000*	*0.0471*
SE	0.9804	0.8291	0.5976	0.9978	0.7876	**0.3227**	0.5200	0.8117	0.9245	0.1358	0.4000	0.9451
	0.0187	*0.1338*	*0.4499*	*0.0014*	*0.1213*	*0.3496*	*0.4057*	*0.1655*	*0.0358*	*0.1203*	*0.3741*	*0.0486*
Swin	**0.9976**	**0.9664**	**0.9785**	**0.9984**	**0.8272**	0.3113	**0.9000**	**0.8206**	**0.9661**	**0.5800**	**1.0000**	**0.9647**
	0.0010	*0.0213*	*0.0139*	*0.0010*	*0.0135*	*0.0172*	*0.0000*	*0.0147*	*0.0184*	*0.2135*	*0.0000*	*0.0192*

Acc and P are short for Accuracy and Precision, respectively. The bold numbers are the best values in the column. The number of cells in each singleton case varies from one to ten. The upper and lower values of each model's row are the mean and standard deviation, respectively.

experiment, our system completes a detection on a singleton case with ten cells within less than 100 seconds, which is one-quarter of the time needed by the semi-automatic method.

References

[1] S. E. Antonarakis. Down syndrome and the complexity of genome dosage imbalance. *Nature Reviews Genetics*, 18(3):147–163, 2017.

[2] B. Feng et al. Down syndrome prediction/screening model based on deep learning and illumina genotyping array. In *Biological Ontologies and Knowledge Bases Workshop at IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)*, IEEE International Conference on Bioinformatics and Biomedicine-BIBM, pages 347–352, 2017.

[3] X. N. Mou, Y. B. Wu, H. H. Cao, Q. Z. Meng, Q. H. Wang, C. C. Sun, S. S. Hu, Y. Ma, and H. Zhang. Generation of disease-specific induced pluripotent stem cells from patients with different karyotypes of down syndrome. *Stem Cell Research & Therapy*, 3, 2012.

[4] H. Cuckle and R. Maymon. Development of prenatal screening-a historical overview. *Seminars in Perinatology*, 40(1):12–22, 2016.

[5] A. C. Neocleous, K. H. Nicolaides, and C. N. Schizas. First trimester noninvasive prenatal diagnosis: A computational intelligence approach. *Ieee Journal of Biomedical and Health Informatics*, 20(5), 2016.

[6] K. H. Nicolaides, K. Spencer, K. Avgidou, S. Faiola, and O. Falcon. Multicenter study of first-trimester screening for trisomy 21 in 75,821 pregnancies: results and estimation of the potential impact of individual risk-orientated two-stage first-trimester screening. *Ultrasound in Obstetrics & Gynecology*, 25(3):221–226, 2005.

[7] D. Somasundaram and V. R. Vijay Kumar. Separation of overlapped chromosomes and pairing of similar chromosomes for karyotyping analysis. *Measurement*, 48:274–281, FEB 2014.

[8] Morag Park. What is a chromosome? *Journal of Pathology*, 163(3):185–189, 1991.

[9] W. C. Schwartzkopf, A. C. Bovik, and B. L. Evans. Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images. *IEEE Transactions on Medical Imaging*, 24(12):1593–1610, 2005.

[10] Enrico Grisan, Enea Poletti, and Alfredo Ruggeri. Automatic segmentation and disentangling of chromosomes in q-band prometaphase images. *IEEE Transactions on Information Technology in Biomedicine*, 13(4):575–581, 2009.

[11] C. Wang, L. Yu, X. Zhu, J. Su, and F. Ma. Extended resnet and label feature vector based chromosome classification. *IEEE Access*, 8:201098–201108, 2020.

[12] Bani Bandana Ganguly, Debasis Banerjee, and Mohan B. Agarwal. Impact of chromosome alterations, genetic mutations and clonal hematopoiesis of indeterminate potential (chip) on the classification and risk stratification of mds. *Blood Cells, Molecules and Diseases*, 69:90–100, Mar 2018.

[13] K. Sundberg, J. Bang, S. SmidtJensen, V. Brocks, C. Lundsteen, J. Parner, N. Keiding, and J. Philip. Randomised study of risk of fetal loss related to early amniocentesis versus chorionic villus sampling. *Lancet*, 350(9079):697–703, 1997.

[14] N. L. Chia. A comprehensive set of idiograms representing all interpretive levels of resolution: Iscn (2009). *Cytogenetic and Genome Research*, 125(2):162–164, 2009.

[15] Jean McGowan-Jordan, Ros J Hastings, and Sarah Moore. *ISCN 2020: An International System for Human Cytogenomic Nomenclature (2020). Reprint Of: Cytogenetic and Genome Research 2020, Vol. 160, No. 7-8.* Karger, S, 2020.

[16] Z. D. Hao, D. K. Lv, Y. Ge, J. S. Shi, D. Weijers, G. C. Yu, and J. H. Chen. Rideogram: drawing svg graphics to visualize and map genome-wide data on the idiograms. *Peerj Computer Science*, 2020.

[17] Y.L. Qin et al. Varifocal-net: A chromosome classification approach using deep convolutional networks. *IEEE Transactions on Medical Imaging*, 38(11):2569–2581, 2019.

[18] C.C Lin et al. A novel chromosome cluster types identification method using resnext wsl model. *Medical Image Analysis*, 69(101943), 2021.

[19] F. Abid and L. Hamami. A survey of neural network based automated systems for human chromosome classification. *Artificial Intelligence Review*, 49(1):41–56, Jan 2018.

[20] B Lerner. Toward a completely automatic neural-network-based human chromosome analysis. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, 28(4):544–52, 1998.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, volume 30, 2017.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[24] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[25] Enea Poletti, Enrico Grisan, and Alfredo Ruggeri. Automatic classification of chromosomes in q-band images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1911–1914, 2008.

[26] E. Pardo, J. M. T. Morgado, and N. Malpica. Semantic segmentation of mfish images using convolutional networks. *Cytometry Part A*, 93A(6):620–627, 2018.

[27] E. Grisan, E. Poletti, and A. Ruggeri. An improved segmentation of chromosomes in q-band prometaphase images using a region based level set. In *World Congress on Medical Physics and Biomedical Engineering*, volume 25, pages 748–751, 2010.

[28] E. Poletti, F. Zappelli, A. Ruggeri, and E. Grisan. A review of thresholding strategies applied to human chromosome segmentation. *Computer Methods and Programs in Biomedicine*, 108(2):679–688, 2012.

[29] Petros Karvelis, Aristidis Likas, and Dimitrios I. Fotiadis. Identifying touching and overlapping chromosomes using the watershed transform and gradient paths. *Pattern Recognition Letters*, 31(16):2474–2488, DEC 1 2010.

[30] Tanvi Arora and Renu Dhir. A variable region scalable fitting energy approach for human metaspread chromosome image segmentation. *Multimedia Tools and Applications*, 78(7):9383–9404, APR 2019.

[31] L. Xiao et al. Deepacev2: Automated chromosome enumeration in metaphase cell images using deep convolutional neural networks. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020.

[32] M. S. Al-Kharraz, L. A. Elrefaei, and M. A. Fadel. Automated system for chromosome karyotyping to recognize the most common numerical abnormalities using deep learning. *IEEE Access*, 8:157727–157747, 2020.

[33] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[34] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[35] W. H. Wang, E. Z. Xie, X. Li, D. P. Fan, K. T. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[37] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016.

[39] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021.

[41] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals, 2021.

[42] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[43] F. Z. Zhuang, Z. Y. Qi, K. Y. Duan, D. B. Xi, Y. C. Zhu, H. S. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the Ieee*, 109(1):43–76, 2021.

[44] 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, 2009.

[45] L. Shen J. Hu and G. Sun. Squeeze-and-excitation networks. In *31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference on Computer Vision and Pattern Recognition, pages 7132–7141, 2018.

[46] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54(1):137–178, JAN 2021.

[47] R. Padilla, S. L. Netto, and E. A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020.

[48] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy, Oct 2017.

[49] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.