Varifocal-Net: A Chromosome Classification Approach Using Deep Convolutional Networks

Yulei Qin, *Student Member, IEEE*, Juan Wen, Hao Zheng, Xiaolin Huang, *Senior Member, IEEE*, Jie Yang, Ning Song, Yue-Min Zhu, Lingqian Wu, and Guang-Zhong Yang, *Fellow, IEEE*

***Abstract*— Chromosome classification is critical for karyotyping in abnormality diagnosis. To expedite the diagnosis, we present a novel method named Varifocal-Net for simultaneous classification of chromosome's type and polarity using deep convolutional networks. The approach consists of one global-scale network (G-Net) and one local-scale network (L-Net). It follows three stages. The first stage is to learn both global and local features. We extract global features and detect finer local regions via the G-Net. By proposing a varifocal mechanism, we zoom into local parts and extract local features via the L-Net. Residual learning and multi-task learning strategies are utilized to promote high-level feature extraction. The detection of discriminative local parts is fulfilled by a localization subnet of the G-Net, whose training process involves both supervised and weakly supervised learning. The second stage is to build two multi-layer perceptron classifiers that exploit features of both two scales to boost classification performance. The third stage is to introduce a dispatch strategy of assigning each chromosome to a type within each patient case, by utilizing the domain knowledge of karyotyping. The evaluation results from 1909 karyotyping cases showed that the proposed Varifocal-Net achieved the highest accuracy per patient case (%) of 99.2 for both type and polarity tasks. It outperformed state-of-the-art methods, demonstrating the effectiveness of our varifocal mechanism, multi-scale feature ensemble, and dispatch strategy. The proposed method has been applied to assist practical karyotype diagnosis.**

***Index Terms*— Chromosome classification, varifocal mechanism, feature ensemble, convolutional networks, dispatch strategy.**

I. INTRODUCTION

CHROMOSOME anomalies, including numerical and structural abnormalities, are responsible for several genetic diseases such as leukemia [1]. Numerical abnormalities arise from the gain or loss of an entire chromosome, which constitute a great proportion of abnormalities [2]. Structural abnormalities result from the breakage and reunion of chromosome segments. In clinical practice, an important procedure for chromosome diagnosis is karyotyping, which is carried out on microscopic images of a single cell [3]. The karyotyping requires first using staining techniques on each cell to obtain stained meta-phase chromosomes. These chromosomes are classified by operators and sorted into 22 pairs of autosomes and 1 pair of sex chromosomes (XX or XY) in the karyotyping map. Then, doctors analyze such map for diagnosis. The karyotyping can be classified into two main categories by the used staining technique: Giemsa karyotyping using Giemsa staining and fluorescent karyotyping using fluorescent staining. If fluorescent staining is employed together with deconvolution of fluorescence signals, chromosomes of different types will be dyed with different colors for fluorescent karyotyping (e.g., SKY [4] and M-FISH [5]). If Giemsa staining is adopted, banding patterns that appear alternatively darker and lighter gray-levels (e.g., G-bands) will be produced for Giemsa karyotyping. Although fluorescent karyotyping is easy for operators to distinguish chromosomes by color, its inherent limitations (e.g., difficulty of detecting all chromosomal abnormalities, impermanent preservation of fluorescence signals, prohibitive cost, controversial reliability of probe hybridization, and unavailability of various probes and clinical samples) make it inappropriate as a first-tier screening tool for examinations [6]–[8]. In contrast, Giemsa karyotyping can detect nearly all abnormalities with a single low-cost test, making it preferred in practice compared to fluorescent karyotyping. A typical karyotyping result map from Giemsa-stained chromosomes is shown in Fig. 1.

The process of karyotyping demands meticulous efforts from well-trained operators. To reduce the burden of

Manuscript received February 23, 2019; accepted March 14, 2019. Date of publication March 19, 2019; date of current version October 25, 2019. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61603248, Grant 61572315, Grant 6151101179, and Grant 81771599, in part by National Key R&D Program of China under Grant 2017YFC1001802, in part by the 863 Plan of China under Grant 2015AA042308, in part by the 973 Plan of China under Grant 2015CB856004, and in part by the 1000-Talent Plan (Young Program) and Committee of Science and Technology, Shanghai, China under Grant 17JC1403000. *(Yulei Qin and Juan Wen contributed equally to this work.) (Corresponding authors: Jie Yang; Ning Song; Lingqian Wu.)*

Y. Qin, H. Zheng, X. Huang, and J. Yang are with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jieyang@sjtu.edu.cn).

J. Wen and L. Wu are with the Center for Medical Genetics, School of Life Sciences, Central South University, Changsha 410078, China (e-mail: wulingqian@sklmg.edu.cn).

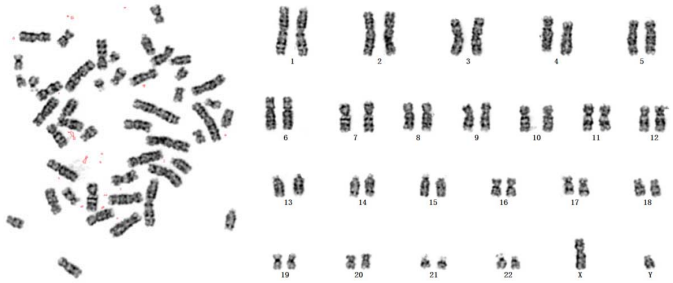
N. Song is with the Shanghai Key Laboratory of Reproductive Medicine, School of Medicine, Shanghai Jiao Tong University, Shanghai 200025, China, and also with the Diagens-Hangzhou, Hangzhou 311121, China (e-mail: ningsong@shsmu.edu.cn).

Y.-M. Zhu is with the University of Lyon, INSA Lyon, CNRS, INSERM, CREATIS UMR 5220, U1206, F-69621 Lyon, France.

G.-Z. Yang is with the Hamlyn Centre for Robotic Surgery, Imperial College London, London SW72AZ, U.K.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2905841



(a) (b)

Fig. 1. (a) A Giemsa-stained microscopic image of male chromosomes for one case. (b) The karyotyping result map of (a) is formed of the paired and ordered chromosomes (22 pairs of autosomes and 1 pair of sex chromosomes XY).

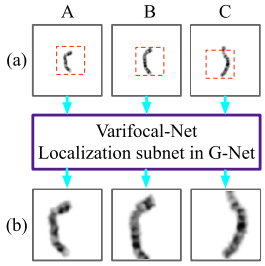


Fig. 2. The focus is varied from global to local. Given chromosome images (A, B, C), the localization subnet detects their finer regions to crop and magnify. (a) The original chromosome images. (b) The local parts after zooming in.

karyotyping, many automated classification methods have been developed for analyzing meta-phase chromosomes [9]–[17]. In general, such methods consist of three steps. The first is to preprocess the chromosome image, which usually involves skeletonization algorithms to compute the medial axis of each chromosome in the image. The second is to extract features along each computed axis. The third step is to build classifiers (e.g., multi-layer perceptron (MLP) and support vector machine (SVM)) to estimate chromosome's type based on the extracted features.

Traditional classification methods mainly rely on geometrical features (e.g., a chromosome's length, centromere position, and banding pattern features). Lerner *et al.* [9] first proposed two approaches of computing medial axis transform (MAT) to detect medial axes of chromosomes. Then, intensity-based features and centromeric indexes were fed into an MLP network for classification. Ming and Tian [10] computed medial axes using a middle point algorithm. They extracted banding patterns by average intensity, gradient, and shape profiles and adopted an MLP classifier. Markou *et al.* [11] proposed a robust method to first extract medial axes using a thinning algorithm. Bifurcations of the axis were removed iteratively via a pixel-neighborhood-based pruning algorithm. Then, the axis was smoothed and extended, with the band-profile features extracted along it. An SVM classifier was finally adopted for type classification. Several other methods targeted at precise detection of the medial axis and centromere location [18]–[21], providing a foundation for accurate chromosome classification.

With the advent of deep learning, researchers tended to employ convolutional neural networks (CNNs) for feature extraction in classification tasks [22]–[30]. Three methods were reported on using deep learning techniques in chromosome studies. Sharma *et al.* [15] proposed a CNN-based method for classification. Bent chromosomes were first straightened by cropping and stitching, and then normalized by length. The accuracy of classification was 86.7% for such preprocessed chromosomes. Gupta *et al.* [16] developed a classification method based on the Siamese Network. Chromosomes were first straightened using two proposed approaches and then fed into the Siamese Network for high-level feature embeddings. An MLP classifier exploited such embeddings for classification and an average accuracy of 85.6% was achieved. Very recently, Wu *et al.* [17] proposed a VGG-Net-D-based approach for category classification. Due to inadequate labeled data, they adopted generative adversarial network (GAN) to generate samples as data augmentation. Their performance was far below requirement for clinical application, with an average precision of 63.5% achieved.

Although many microscopes are nowadays equipped with chromosome classification systems (e.g., CytoVision [31]–[33], Ikaros [34], and ASI HiBand [35]), users still have to manually drag each chromosome image and drop it to the target position in the practical karyotyping process due to their poor performance. Research studies reveal that the challenges of chromosome classification mainly lie in the following aspects: 1) Chromosomes are often curved and bent due to their non-rigid nature, making it difficult to accurately extract their medial axes. Hence, errors accumulate in the process of straightening and feature computation along such axes, leading to an accuracy drop. 2) Even for the chromosomes of the same class, they vary slightly from person to person in terms of local details. The generalizability and performance of traditional methods, which depend on manually designed features, may degrade for clinical applications. 3) The chromosome polarity, which reflects whether a chromosome's q-arm (long arm) is downward or upward, is often not considered in previous work. However, it is important to decide the chromosome's orientation in the process of repositioning for the karyotyping map generation. All the q-arms should stay downward.

To tackle the above challenges, we propose a novel CNN-based approach for chromosome classification. Its name, Varifocal-Net, highlights the capacity to zoom into local regions automatically. It has one global-scale network (G-Net) and one local-scale network (L-Net). We extract global features and pinpoint specific local regions via the G-Net. The view is changed (see Fig. 2) as our Varifocal-Net zooms into the discriminative region of a chromosome. Local features are extracted from such local parts via the L-Net. At first glance, such a global-to-local idea resembles the concept of multi-scale CNNs used in cellular image analysis [36]–[39] and other vision tasks [40]–[42]. However, unlike previous multi-scale methods, our approach learns multi-scale information in the global-to-local mechanism. It locates the discriminative local region and extracts the features of the two scales through two

independent networks. The proposed Varifocal-Net comprises three stages. The first stage is to learn effective feature representations at both global and local scales. The global-scale representations mainly concern overall information such as the chromosome's length, shape, and size, which determines its type on a coarse-grained level. The local-scale representations depict details such as texture patterns of local parts, which facilitate discrimination among chromosomes on a fine-grained level. The second stage is to build two MLP classifiers to leverage features of both two scales for prediction of type and polarity, respectively. The third stage is to introduce a dispatch strategy for type assignment within each patient case. To validate the effectiveness and generalizability of our approach, we construct a large dataset containing 1909 karyotyping cases. Extensive experiments on the dataset corroborate that the Varifocal-Net achieved better performance than state-of-the-art methods. Our contributions can be summarized as follows:

- Inspired by the zoom capability of cameras, we propose the Varifocal-Net to address the challenges of chromosome classification. We extract global-scale features from the whole image and local-scale features from the local region selected by our varifocal mechanism. Residual learning and multi-task learning strategies are utilized to promote effective feature learning. The detection of discriminative local parts is fulfilled via a localization subnet whose training involves both supervised and weakly-supervised learning.
- We utilize the concatenated features from both global and local scales to predict type and polarity simultaneously, thereby combining the knowledge acquired at two scales. To our best knowledge, this represents the first attempt to take multi-scale feature ensemble into account in chromosome studies.
- We propose a dispatch strategy to assign each chromosome to a type based on its predicted probabilities. Both the maximum likelihood criterion and possible abnormality situations are taken into account to enable the strategy suitable for clinical settings.
- We evaluate the proposed approach on a large dataset. It demonstrates its superior performance compared with state-of-the-art methods. The end-to-end manner of classification sidesteps the problem of inaccurate medial axis extraction and chromosome straightening.
- The Varifocal-Net has been put into clinical practice for chromosome classification. For each patient, it accurately classifies both abnormal and healthy chromosomes and diagnoses numerical abnormalities if the number of classified chromosomes is irregular.

The paper is structured as follows: In Section II, we describe the proposed method. In Section III, we provide experiments and results. Section IV discusses our findings, followed by the conclusion in Section V.

II. Methods

The proposed Varifocal-Net is composed of three stages: a) Global-scale and local-scale feature learning by optimizing the Varifocal-Net in an alternative way; b) Classification of type and polarity via MLP classifiers utilizing the fused features; c) Assignment of chromosomes' types with the proposed dispatch strategy. Original chromosome images are separated manually by cytogeneticists from captured microscopic images. They are preprocessed as discussed in Sec. III-B and taken as inputs to the G-Net in the first stage. The G-Net contains deep CNNs, one classification subnet, and one localization subnet, as shown in Fig. 3. Global-scale features are extracted via the CNNs, which are optimized by the loss function of the classification subnet. After the CNNs and classification subnet converge, we pre-train the localization subnet to output initial coordinates for local region detection. Then, with local parts cropped and rescaled, we optimize the L-Net and the localization subnet of the G-Net alternatively. In the second stage, with the fused two-scale features, we build two MLP classifiers to predict chromosome's type and polarity, respectively. The schematic representations of the first stage and the second stage of our Varifocal-Net are illustrated in Fig. 3 and Fig. 6, respectively. For each chromosome within one patient case, a dispatch strategy is employed in the third stage to assign it to a certain type based on its predicted probabilities.

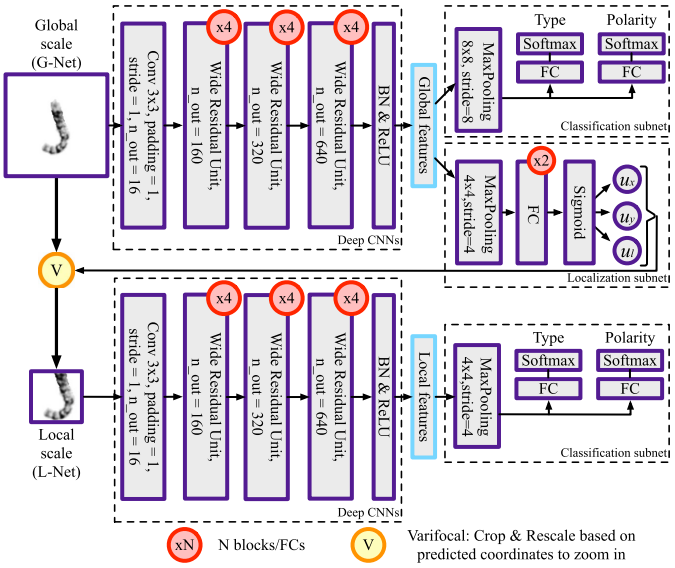


Fig. 3. The first stage of the proposed Varifocal-Net: global-scale and local-scale feature extraction via the G-Net and the L-Net, respectively.

A. Stage 1: Global-Scale and Local-Scale Feature Learning

1) Feature Extraction With Residual Learning: The architecture of deep CNNs for feature extraction is the same for both the G-Net and the L-Net. Inspired by the success of ResNet [26], [43], we adopt wide residual blocks to introduce residual learning. Such CNNs consist of one convolution layer (Conv), three residual blocks, one batch normalization layer (BN), and one rectified linear unit (ReLU). Each residual block has four residual units as illustrated in Fig. 4, with the first unit

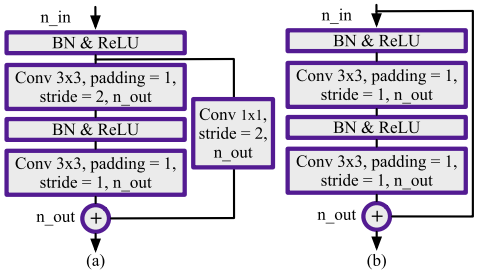


Fig. 4. Wide residual unit. n_in and n_out stand for number of input and output feature channels, respectively. (a) if n_in $\neq$ n_out. (b) if n_in = n_out.

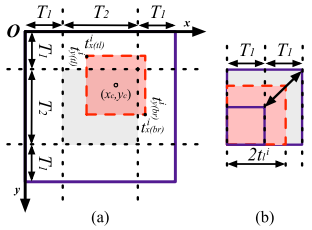


Fig. 5. The diagram of parameterizations for the sample x_i. (a) The red box is the predicted local region and the gray background square is the area where the box's center pixel (x_c, y_c) can be located. (b) The side length of the predicted box $(2t_l^i)$ is restricted, ranging from T_1 to $2T_1$.

increasing the number of channels and downsampling features through strided convolution.

2) Multi-Task Learning With Weighted Classification Loss: Since the tasks of type and polarity classification are correlated, we adopt multi-task learning to take inner relation between these tasks into consideration. It improves the effectiveness of feature extraction through a shared representation of CNNs [44]. In the classification subnet, a max-pooling layer is followed by two fully-connected (FC) layers respectively to predict type and polarity. The FC layers map the feature vector to the probability vectors of 24 dimensions (for the type task) and 2 dimensions (for the polarity task). We train the deep CNNs in the G-Net and the L-Net independently by minimizing a weighted loss of the classification subnet. For the type task, given a set of N training triplets $\{(x_i, y_i^t, y_i^p)\}_{i=1,2,...,N}$, the cross-entropy loss between the output vector $\mathbf{O^t}$ and the target vector $\mathbf{Y^t}$ is given by:

$$\mathcal{L}_t(\mathbf{O^t}, \mathbf{Y^t}) = \sum_{i=1}^{N} -log(\frac{exp(o_i^t[y_i^t])}{\sum_{j=1}^{24} exp(o_i^t[j])}), \tag{1}$$

where o_i^t and y_i^t denote the output probability vector and the target type for the sample x_i, respectively. Note that here we combine the softmax function and the standard cross-entropy function into one formula. Similarly, the polarity classification loss between the predicted vector $\mathbf{O^p}$ and the target vector $\mathbf{Y^p}$ is defined as:

$$\mathcal{L}_p(\mathbf{O^p}, \mathbf{Y^p}) = \sum_{i=1}^{N} -log(\frac{exp(o_i^p[y_i^p])}{\sum_{j=1}^{2} exp(o_i^p[j])}), \tag{2}$$

where o_i^p and y_i^p stand for the probability vector and the target polarity, respectively. The total multi-task loss is given by:

$$\mathcal{L}_{cls}(\mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p}) = \mathcal{L}_t(\mathbf{O^t}, \mathbf{Y^t}) + \lambda \mathcal{L}_p(\mathbf{O^p}, \mathbf{Y^p}), \tag{3}$$

in which λ is a weight controlling the balance between the two loss terms. We place more emphasis on the type task, thus setting $\lambda = 0.5$ in our experiments.

3) Varifocal Mechanism: Previous work on chromosome classification takes no advantage of multi-scale feature learning and fusing. These methods do not detect specific finer parts for detail description (e.g., nuance of banding's number, width, and intensity among similar chromosomes). Motivated by the success of region proposal network (RPN) [45], [46] and attention proposal network (APN) [29], we propose a varifocal mechanism that zooms into local regions of chromosomes automatically for finer feature extraction. Given a chromosome sample x_i, it first predicts the position and size of a local region box via the localization subnet, which is sequentially composed of a max-pooling layer, two FC layers, and a sigmoid layer. The square box prediction is expressed as:

$$(u_x^i, u_y^i, u_l^i) = f(\mathbf{W_c} * x_i), \tag{4}$$

where $\mathbf{W_c}$ and $*$ denote all parameters of deep CNNs and their related operations (e.g., Conv, BN, and ReLU), respectively. $\mathbf{W_c} * x_i$ gives the global feature of x_i and $f(\cdot)$ represents the proposed localization subnet. The variables u_x^i and u_y^i denote the relative coordinates of the box's center (x_c, y_c) and u_l^i is the relative length of the half of its side. All these variables range from 0 to 1. Assuming the top-left corner of x_i as the origin of the global pixel coordinate system where x-axis starts from left to right and y-axis from top to bottom, we adopt the parameterizations of the top-left (tl) and bottom-right (br) pixels of the region box as follows:

$$\begin{aligned} t_{x(tl)}^i &= T_1 + u_x^i \cdot T_2 - t_l^i, \quad t_{x(br)}^i = T_1 + u_x^i \cdot T_2 + t_l^i, \\ t_{y(tl)}^i &= T_1 + u_y^i \cdot T_2 - t_l^i, \quad t_{y(br)}^i = T_1 + u_y^i \cdot T_2 + t_l^i, \\ t_l^i &= u_l^i \cdot T_1/2 + T_1/2, \end{aligned} \tag{5}$$

where T_1, T_2, and t_l^i denote the minimum margin, maximum shift, and half of the side length, respectively. Fig. 5 illustrates these parameterizations. Note that here we restrict the position and size of the predicted local region for two reasons. First, the predicted region should focus on a discriminative part of the chromosome, which is in the center of the image. Second, the region cannot exceed the image boundary and its size should be moderate to effectively capture local features. In our implementation, we set $T_2 = 2T_1$ empirically because it forces the localization subnet to focus on the central region.

Once a local region is predicted, the focus is moved onto it by cropping and rescaling. The cropping operation is implemented using a variant of two-dimensional (2-D) boxcar function [29] as an approximation. Given the coordinate tuple $(t_{x(tl)}^i, t_{y(tl)}^i, t_{x(br)}^i, t_{y(br)}^i)$, we use the boxcar function to generate a region mask and multiply it with the original image in an element-wise manner. It is mathematically expressed as:

$$\begin{aligned} x_i^{loc} &= x_i \odot boxcar(t_x^i, t_y^i, t_l^i), \\ boxcar(t_x^i, t_y^i, t_l^i) &= (H(x - t_{x(tl)}^i) - H(x - t_{x(br)}^i)) \\ &\quad \cdot (H(y - t_{y(tl)}^i) - H(y - t_{y(br)}^i)) \end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication and x_i^{loc} stands for the cropped local part. The 2-D $boxcar(t_x^i, t_y^i, t_l^i)$ function serves as a mask and $H(x)$ is the Heaviside step function. Note that the derivative of $H(x)$ is infinite at $x = 0$. Since its derivative is required in back-propagation, we use the logistic function as a smooth analytic approximation for $H(x)$ in experiments, which is computed by:

$$H(x) = \frac{1}{1 + e^{-kx}}, \quad k > 0 \tag{7}$$

in which a larger k (e.g., $k = 10$) leads to a sharper change at $x = 0$. The multiplication with $boxcar(t_x^i, t_y^i, t_l^i)$ will mask out the target local region by keeping the value of pixels inside the region almost unchanged and that of others close to zero. Then, we crop the target region in x_i^{loc} and rescale it to a unified size via bilinear interpolation, which makes it easier for both algorithm implementation and finer feature extraction in the L-Net. So far, the Varifocal-Net has zoomed into a particular local part. Note that in the forward process, the local region is cropped directly by indexed slicing. In the backward propagation process, since the cropping operation is not derivative, the boxcar function is used to approximate it and provide necessary gradient for proper parameter optimization. Detailed analytical derivations are presented in Sec. II-A.5.

4) Loss Function of the Localization Subnet: With definitions of the localization subnet $f(\cdot)$, we adopt both supervised and weakly-supervised learning to optimize it. The supervised method is employed in pre-training to initialize the parameters of $f(\cdot)$. For such pre-training, we assign the ground-truth coordinates $(u_x^{i*}, u_y^{i*}, u_l^{i*})$ for the sample x_i as follows: 1) The locations u_x^{i*} and u_y^{i*} are set to 0.5 since a chromosome is centered in the image. 2) Based on u_x^{i*} and u_y^{i*}, the smallest region that covers the whole chromosome is calculated and u_l^{i*} is computed accordingly. The lower bound of u_l^{i*} is 0 and if the width or height of chromosome exceeds $2T_1$, u_l^{i*} will be set to 1. Given a set of N sample pairs $\{(x_i, u_x^{i*}, u_y^{i*}, u_l^{i*})\}_{i=1,2,\ldots,N}$, our loss function for supervised learning is defined as:

$$\begin{aligned} \mathcal{L}_u(\mathbf{U}, \mathbf{U}^*) &= \sum_{i=1}^{N} \sum_{\gamma \in \{x,y,l\}} \text{smooth}_{L_1}(u_\gamma^i - u_\gamma^{i*}), \\ \text{smooth}_{L_1}(x) &= \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \end{aligned} \tag{8}$$

where $\mathbf{U}$ and $\mathbf{U}^*$ denote the vector of the predicted coordinates and their ground-truth labels, respectively. The robust smooth_{L_1} loss [45] is used to directly train the localization subnet to output initial local region coordinates. It is less sensitive to outliers than L_2 loss and smoother near zero compared to the standard L_1 norm. Its gradient is uniquely defined at zero point.

While the weakly-supervised method is aimed at improving the classification performance of the L-Net by optimizing the $f(\cdot)$ for finer part localization, we keep all parameters of the L-Net unchanged and only fine-tune the localization subnet by minimizing the multi-task loss (3) of the L-Net. Without ground-truth coordinates provided, the subnet $f(\cdot)$ autonomously learns to locate discriminative parts, making the extracted features meaningful. Thus, the total loss is given by:

$$\begin{aligned} \mathcal{L}_{loc}(\mathbf{U}, \mathbf{U}^*, \mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p}) &= \mathcal{L}_u(\mathbf{U}, \mathbf{U}^*) \\ &+ \mathcal{L}_{cls}(\mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p}). \end{aligned} \tag{9}$$

Here, the subnet is only pre-trained once by minimizing $\mathcal{L}_u(\mathbf{U}, \mathbf{U}^*)$. Then, its optimization process is dominant by weakly-supervised learning. The training details of our proposed Varifocal-Net will be introduced in Sec. II-C.

5) Back-Propagation Through Boxcar Function: We adopt the boxcar function for localization because it provides analytical representations between region cropping and the predicted relative coordinates (u_x^i, u_y^i, u_l^i), which is indispensable for parameter update in back-propagation. When optimizing $\mathcal{L}_{cls}(\mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p})$ to train the localization subnet, gradients back-propagate through the boxcar function. For one single image x_i, we designate the gradients that back-propagate to the input layer of the L-Net as $\mathbf{G}_{top}$. The partial derivatives of the loss to coordinates are then given by:

$$\begin{aligned} \frac{\partial \mathcal{L}_{cls}(\mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p})}{\partial u_\gamma^i} &\propto \mathbf{G}_{top} \odot \frac{\partial boxcar(t_x^i, t_y^i, t_l^i)}{\partial t_\gamma^i} \cdot \frac{\partial t_\gamma^i}{\partial u_\gamma^i}, \\ \frac{\partial t_x^i}{\partial u_x^i} = \frac{\partial t_y^i}{\partial u_y^i} &= T_2, \ \frac{\partial t_l^i}{\partial u_l^i} = T_1/2, \ \gamma \in \{x, y, l\}, \end{aligned} \tag{10}$$

where $\odot$ denotes element-wise multiplication. Hence, the derivatives of $boxcar(t_x^i, t_y^i, t_l^i)$ with respect to t_x^i, t_y^i and t_l^i largely influence the moving direction and size of the local region box. Note that in the context of minimizing our loss, it holds true for $\forall \gamma \in \{x, y, l\}$ that t_γ^i increases when $\frac{\partial \mathcal{L}_{cls}(\mathbf{O^t}, \mathbf{Y^t}, \mathbf{O^p}, \mathbf{Y^p})}{\partial u_\gamma^i} < 0$ and decreases otherwise. To achieve a consistent optimization direction, we follow [29] to calculate the negative squared norm of derivatives $\mathbf{G}_{top}$ and compute the $boxcar(t_x^i, t_y^i, t_l^i)$'s partial derivatives explicitly in the back-propagation process.

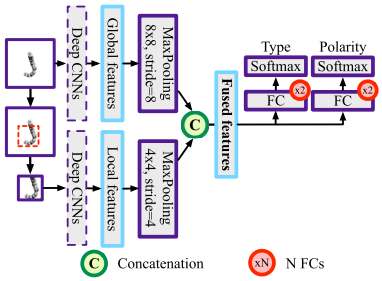


Fig. 6. The second stage of the proposed Varifocal-Net: chromosome classification using fused features from both global and local scales.

B. Stage 2: Classification Based on the Fused Feature

Once both the G-Net and the L-Net are optimized, global-scale and local-scale features can be extracted via deep CNNs. To make full use of these two representations, it is reasonable to concatenate them into a feature ensemble. We build two MLP classifiers (see Fig. 6) to learn the mapping from the

fused features to classification probabilities of type and polarity, respectively. Each classifier consists of two FC layers and one Softmax layer. With the trained classifiers, the proposed Varifocal-Net simultaneously predicts chromosome's type and polarity in an end-to-end manner.

C. Four-Step Training Strategy

In this paper, we adopt a four-step optimization technique to alternatively train the network. In the first step, we initialize deep CNNs of the G-Net and L-Net via He's method [47]. In the second step, we train deep CNNs and the classification subnet in the G-Net until convergence. At this point, the localization subnet and the L-Net are not optimized. In the third step, we prepare all the ground-truth coordinates of local region boxes and only pre-train the localization subnet once. Finally, we train the L-Net and the localization subnet alternatively in the fourth step. Keeping the parameters of the localization subnet fixed, we optimize the L-Net by minimizing our multi-task loss. Then we fix the parameters of the L-Net and fine-tune the localization subnet alone. Such alternative training can be run for iterations until there is no further error loss decrease.

D. Stage 3: Type Assignment Using Dispatch Strategy

In karyotyping practice, the classification of chromosome's type is conducted within each patient case. Therefore the classification can also be viewed as dispatching each chromosome to a certain type. This led us to propose a dispatch strategy for type assignment in the third stage. The design of the dispatch strategy follows two simple rules about karyotyping's domain knowledge [48]:

- Each healthy patient has 46 chromosomes for 23 classes (female) or 24 classes (male).
- For unhealthy patient, the number of each type falls between 1 and 3 (e.g., monosomy 21 and trisomy 21) except extremely rare cases. Type Y has less than 3 chromosomes.

Considering both the maximum likelihood criterion and possible abnormality situations, we dispatch chromosomes twice. Given the predicted probabilities from the second stage of the Varifocal-Net, the first-time dispatch is to assign each chromosome to the type having the highest probability. The second-time dispatch is to check and compare the probabilities of different chromosomes that are assigned to the same type. The confidence threshold th is designed to filter out uncertain assignments. The dispatch strategy is described in details in Alg. 1. Note that it is not used for polarity prediction because polarity only involves 2 classes (q-arm upward or downward).

Algorithm 1 Dispatch strategy for chromosome's type

Input: N chromosomes; the probabilities of 24 types P_i for the i-th chromosome (P_{ij} stands for its probability of being type j, $i = 1, 2, ..., N$, $j = 1, 2, ..., 24$); confidence threshold th.

Output: The set of chromosomes assigned to type k ($O_k, k = 1, 2, ..., 24$); possible abnormal warnings.

$T_k = \varnothing$, $O_k = \varnothing$, $\forall k \in \{1, 2, ..., 24\}$.
for each $i \in \{1, 2, ..., N\}$ **do**
  Compute the most probable type $j^* = \arg\max_j P_{ij}$ and dispatch the i-th chromosome to type j^* by $T_{j^*} = T_{j^*} \cup \{i\}$;
end for
for each $k \in \{1, 2, ..., 24\}$ **do**
  $S = 1$ if $k = 24$, otherwise $S = 2$;
  if $|T_k| > S$ **then**
    Sort each element in T_k based on its probability. From T_k, choose $S + 1$ elements ($Q_k = \{i^1, ..., i^{S+1}\}$) with the highest probability if $P_{ik} > th, \forall i \in Q_k$, otherwise choose only S elements ($Q_k = \{i^1, ..., i^S\}$);
    $O_k = O_k \cup Q_k$;
    for $i \in T_k \setminus Q_k$ **do**
      Compute the second probable type $j^* = \arg\max_{j, j \neq k} P_{ij}$ and dispatch it to type j^* by $O_{j^*} = O_{j^*} \cup \{i\}$;
    end for
  else
    $O_k = O_k \cup T_k$;
  end if
end for
Print abnormal warnings if $|O_k| \neq 2, \forall k \in \{1, 2, ..., 22\}$ or $|O_{23}| + |O_{24}| \neq 2$;
return $O_k, k = 1, 2, ..., 24$.

III. Experiments and Results

A. Materials

For the experiments conducted in this section, we collected 1909 different patients' karyotyping cases from the Xiangya Hospital of Central South University, China. Each patient case contains one Giemsa stained microscopic image of meta-phase chromosomes. All images are grayscale and sampled with the same resolution, using the Leica's CytoVision System (GSL-120). Each chromosome is of approximate 300-band levels. The datasets contain 1784 karyotyping cases from healthy patients (1061 male and 723 female) and 125 cases from unhealthy patients (73 male and 52 female). The unhealthy cases contain both numerical and structural abnormalities. Each chromosome's type is manually annotated by cytogeneticists in real-world clinical environments. The type of autosomes is labeled from 0 to 21 and the type of sex chromosomes X and Y are denoted as 22 and 23, respectively. The polarity of a chromosome is labeled as 1 if its q-arm is downward and 0 otherwise.

We obtain each individual chromosome image by manually segmenting it from microscopic images. In total, there exist 87831 separated chromosomes. We randomly split both healthy and unhealthy samples into five subsets to perform five-fold cross validation. Each time, four subsets are used for training the model and fine-tuning the hyper-parameters. The remaining one subset is left for testing. Note that the chromosome samples are divided by patient case. All chromosomes

TABLE I
STATISTICS OF THE DATASET. (H: HEALTHY SAMPLES, U: UNHEALTHY SAMPLES.)

Dataset		Case #		Image #		Total image #
		Male	Female	Male	Female	
Total samples	H	1061	723	48806	33258	87831
	U	73	52	3384	2383	

TABLE II
THE FEATURE DIMENSIONS OF THE VARIFOCAL-NET FOR THE FIRST STAGE. (T: TYPE, P: POLARITY, LOC: LOCALIZATION.)

Layer	Dimension				
	G-Net			L-Net	
Input	256×256			128×128	
Deep CNNs	$640 \times 32 \times 32$			$640 \times 16 \times 16$	
Max-pooling	$640 \times 4 \times 4$		$640 \times 8 \times 8$	$640 \times 4 \times 4$	
FC1	24 (T)	2 (P)	1024	24 (T)	2 (P)
FC2	–	–	3 (Loc)	–	–

TABLE III
THE FEATURE DIMENSIONS OF THE VARIFOCAL-NET FOR THE SECOND STAGE. (T: TYPE, P: POLARITY.)

Layer	Dimension	
Input	256×256 (G-Net)	128×128 (L-Net)
Deep CNNs	$640 \times 32 \times 32$	$640 \times 16 \times 16$
Max-pooling	$640 \times 4 \times 4$	$640 \times 4 \times 4$
Concatenation	$640 \times 4 \times 4 \times 2$	
FC1	512	512
FC2	24 (T)	2 (P)

of the same case stay in the same subset. Table I provides the details of our datasets.

B. Implementation Details

The size of images differs from each other and we first padded them with pixels into square images of the same size. The padding value is set as 255 to imitate the background of the original Giemsa stained images. And the size of padded image is 320×320 pixels. Then, we resized the image to 256×256 pixels and normalized all N images as follows:

$$x_i' = (x_i - \mu_i)/\sigma_i, \quad i = 1, 2, ..., N \tag{11}$$

where μ_i and σ_i are the mean value and the standard deviation of the sample x_i, respectively. x_i' denotes the normalized input, which has a zero mean and a unit variance. For local region prediction, the margin T_1 is 64 and the shift range T_2 is 128. The cropped target region was then upsampled to 128×128 pixels as the input to the L-Net. In Table II and Table III, we describe feature dimensions of the proposed Varifocal-Net for the first and the second stages, respectively. For the dispatch strategy, the confidence threshold th was set to 0.9 because we only keep highly-confident chromosomes when possible numerical abnormalities happen.

In the training process, we adopted horizontal flipping and random rotation between $[0°, 45°]$ for data augmentation. The vertical flipping operation was performed to change the polarity label of a chromosome. All modules of the Varifocal-Net were trained from scratch using Adam optimizer [49] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to 0.0001 and it decreased by nine-tenth every 10 epochs. We implemented the proposed Varifocal-Net and other CNN-based methods in Python, with PyTorch framework [50]. All experiments were conducted under a Ubuntu OS workstation with Intel Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 128 GB of RAM, and 4 NVIDIA GTX Titan X GPUs.

C. Evaluation Metrics

The performance of the Varifocal-Net was evaluated by four metrics: the accuracy of all the testing images (Acc.), the average F_1-score over classes of all the testing images (F_1), the average accuracy of the complete karyotyping per patient case (Acc. per Case), and the average accuracy of the complete karyotyping per patient case using the proposed dispatch strategy (Acc. per Case-D). The Acc. is an intuitive measurement defined as the fraction of the testing samples which are correctly classified.

For the computation of F_1-score, we first define the following four criteria to fit the context of multi-class classification:

- True positives (TP_j): images predicted as class j which actually belong to class j
- False positives (FP_j): images predicted as class j which actually do not belong to class j
- False negatives (FN_j): images predicted as class k ($\forall k \neq j$) which actually belong to class j
- True negatives (TN_j): images predicted as class k ($\forall k \neq j$) which actually do not belong to class j

Then, the F_1-score is computed as:

$$\begin{aligned} F_1 &= \frac{1}{N_{cls}} \sum_{j=1}^{N_{cls}} \frac{2 \cdot Precision_j \cdot Recall_j}{Precision_j + Recall_j}, \\ Precision_j &= \frac{TP_j}{TP_j + FP_j}, \\ Recall_j &= \frac{TP_j}{TP_j + FN_j}, \end{aligned} \tag{12}$$

where N_{cls} equals 24 and 2 for type and polarity recognition, respectively.

The accuracy per patient case was adopted to evaluate the performance in clinical settings. It is computed by checking the fraction of the correctly classified samples within each patient case. No dispatch strategy is used for computing Acc. per Case. We only assign each chromosome to the type having the highest predicted probability. For the computation of Acc. per Case-D, the proposed dispatch strategy is employed and accuracy within each case is recalculated for all samples.

The mean value and the standard deviation of these four metrics are provided to assess performance stability. They were calculated based on the results of five-fold cross validation and displayed in percentage.

Furthermore, we also adopted a receiver operating characteristic (ROC) analysis for performance comparison. The ROC curves averaged over all classes were plotted and the area under each curve (AUC) was calculated as well.

TABLE IV

PERFORMANCE OF THE VARIFOCAL-NET (MEAN±STANDARD DEVIATION). THE RESULTS ARE PRESENTED IN TERMS OF FOUR EVALUATION METRICS: AVERAGE F_1-SCORE OF ALL TESTING IMAGES (F_1), ACCURACY OF ALL TESTING IMAGES (ACC.), AVERAGE ACCURACY PER PATIENT CASE (ACC. PER CASE), AND AVERAGE ACCURACY PER PATIENT CASE USING THE PROPOSED DISPATCH STRATEGY (ACC. PER CASE-D). (T: TYPE, P: POLARITY, PET: PER EPOCH TIME, TPI: TIME PER IMAGE.)

Stage	Method	F_1 (%)		Acc. (%)		Acc. per Case (%)		Acc. per Case-D (%)	# Epoch ×PET (s)	Testing TPI (ms)
		T	P	T	P	T	P	T		
1	G-Net	97.5±0.4	99.0±0.1	97.8±0.4	99.0±0.1	97.8±3.8	99.0±1.9	98.2±3.3	30×956.3±1.5	5.7±0.1
	L-Net	98.2±0.5	99.2±0.1	98.4±0.5	99.2±0.1	98.4±2.9	99.2±1.6	98.9±2.5	30×1142.3±2.3	6.8±0.1
2	**Varifocal-Net**	**98.7±0.7**	**99.2±0.3**	**98.9±0.7**	**99.2±0.3**	**98.9±2.3**	**99.2±1.5**	**99.2±2.1**	**20×1150.8±11.3**	**5.9±0.1**

TABLE V

PERFORMANCE OF THE VARIFOCAL-NET FOR EACH CHROMOSOME TYPE (MEAN±STANDARD DEVIATION)

Class (No.)	F_1 (%)	Precision (%)	Recall (%)
1	99.6±0.6	99.5±0.7	99.7±0.5
2	99.3±0.7	98.8±0.9	99.7±0.5
3	99.5±0.6	99.4±0.8	99.6±0.5
4	98.6±1.1	98.4±1.1	98.7±1.1
5	98.6±0.7	98.7±0.7	98.6±0.7
6	99.4±0.6	99.7±0.3	99.2±0.8
7	99.7±0.2	99.7±0.3	99.6±0.3
8	98.9±0.8	98.9±0.9	98.9±0.8
9	98.7±0.4	98.8±0.8	98.6±0.5
10	98.7±0.7	98.7±0.8	98.7±0.7
11	99.6±0.2	99.6±0.3	99.6±0.3
12	99.7±0.2	99.8±0.1	99.6±0.4
13	98.7±0.7	98.8±0.5	98.7±1.0
14	99.0±0.5	99.2±0.6	98.9±0.5
15	98.5±0.8	98.6±0.9	98.4±0.7
16	97.9±1.3	97.9±1.2	97.9±1.4
17	99.3±0.6	99.1±0.8	99.4±0.4
18	98.8±1.1	98.9±0.8	98.7±1.5
19	98.7±0.8	98.6±0.9	98.8±0.8
20	98.4±1.0	98.5±1.1	98.4±0.9
21	98.5±0.5	98.5±0.4	98.6±0.7
22	98.4±0.8	98.3±0.8	98.6±0.9
X	98.3±1.1	98.6±0.9	98.1±1.4
Y	94.3±3.6	95.0±3.5	93.6±3.8

TABLE VI

PERFORMANCE OF THE VARIFOCAL-NET FOR EACH CHROMOSOME POLARITY (MEAN±STANDARD DEVIATION)

Class	F_1 (%)	Precision (%)	Recall (%)
q-arm upward	99.2±0.3	99.1±0.4	99.4±0.1
q-arm downward	99.3±0.3	99.4±0.1	99.1±0.4

TABLE VII

PERFORMANCE OF THE VARIFOCAL-NET FOR POLARITY CLASSIFICATION WITHIN EACH TYPE (MEAN±STANDARD DEVIATION)

Class (No.)	Acc. (%)	Class (No.)	Acc. (%)	Class (No.)	Acc. (%)
1	99.3±0.5	9	99.6±0.1	17	99.3±0.5
2	99.1±0.5	10	99.5±0.2	18	99.5±0.1
3	99.5±0.4	11	99.8±0.2	19	99.3±0.4
4	99.2±0.4	12	99.6±0.2	20	96.2±1.1
5	99.1±0.3	13	99.6±0.4	21	99.3±0.3
6	99.5±0.2	14	99.8±0.2	22	99.5±0.1
7	99.8±0.2	15	98.9±0.5	X	99.2±0.3
8	99.5±0.2	16	99.1±0.4	Y	98.0±0.8

D. Results

This section presents experimental results in three parts. We first provide detailed evaluation results of the proposed Varifocal-Net. Then, a comparison of the proposed method with state-of-the-art methods is given. Finally, we present additional results for analyzing our performance.

1) Evaluation Results: Table IV gives the classification results of the G-Net, L-Net, and the entire Varifocal-Net. The global-scale G-Net achieved the accuracy (%) of 97.8 and 99.0 for type and polarity recognition, respectively. With the localization subnet for finer region detection, the local-scale L-Net reduced classification errors. By utilizing the knowledge learned at two scales, the proposed Varifocal-Net yielded the best performance. The accuracy (%) of type and polarity tasks were boosted to 98.9 and 99.2, respectively. Due to the proposed dispatch strategy, the accuracy of type classification per case is further improved for each method. The proposed Varifocal-Net achieved the averaged Acc. per Case-D (%) of 99.2. Though the total training time is relatively long, the testing time of the Varifocal-Net is only 5.9ms per sample.

To observe the performance of the Varifocal-Net on each class of chromosomes, Table V and Table VI provide the F_1-score, precision, and recall, which were computed within each category. For type recognition, the proposed method performed worst on Y chromosomes, with only a F_1-score (%) of 94.3 achieved. The evaluation results of classes No. 4, No. 5, No.15, No. 16, No. 20–No. 22, X, and Y are below average. For polarity recognition, the orientation of q-arm was accurately predicted, with the F_1-score of each class above 99%.

Besides, for polarity classification, we also computed the accuracy within each type category to learn the performance difference among chromosome types. Table VII indicates that our prediction is relatively inaccurate for two long types (classes No. 2 and No. 5) and four short types (classes No. 15, No.16, No. 20 and Y).

2) Comparison With the State-of-the-Art: Table VIII provides a comparison of the proposed Varifocal-Net with state-of-the-art methods. The first two methods [15], [16] were proposed specifically for classifying Giemsa stained chromosomes. Both the two existing methods employed CNNs for feature extraction, and they relied on straightening chromosomes for normalization and used small datasets. In contrast, we adopted an end-to-end fashion for prediction. We implemented the two methods and evaluated them using five-fold cross validation. Their performance of type recognition on our large testing set proves the superiority of our method, which surpasses [15] and [16] by nearly 6.7% and 7.5% in average F_1-score, respectively.

TABLE VIII

COMPARISON RESULTS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS (MEAN±STANDARD DEVIATION). THE RESULTS ARE PRESENTED IN TERMS OF FOUR EVALUATION METRICS: AVERAGE F_1-SCORE OF ALL TESTING IMAGES (F_1), ACCURACY OF ALL TESTING IMAGES (ACC.), AVERAGE ACCURACY PER PATIENT CASE (ACC. PER CASE), AND AVERAGE ACCURACY PER PATIENT CASE USING THE PROPOSED DISPATCH STRATEGY (ACC. PER CASE-D). (T: TYPE, P: POLARITY.)

Method	F_1 (%)		Acc. (%)		Acc. per Case (%)		Acc. per Case-D (%)
	T	P	T	P	T	P	T
Sharma *et al.* [15]	92.0±1.6	–	92.6±1.5	–	92.6±7.9	–	93.6±7.4
Gupta *et al.* [16]	91.2±2.3	–	91.8±2.2	–	91.8±9.9	–	92.6±9.5
AlexNet [23]	90.2±1.9	97.1±0.5	90.8±1.8	97.1±0.5	90.8±9.5	97.1±3.9	92.4±9.1
GoogLeNet [24]	95.6±1.6	98.6±0.5	96.0±1.5	98.6±0.5	96.0±6.5	98.6±2.7	96.8±6.2
VGG-Net [25]	96.0±0.7	98.8±0.2	96.3±0.6	98.8±0.2	96.3±5.3	98.8±2.2	97.1±4.9
ResNet [26]	96.6±0.9	98.9±0.2	96.9±0.9	98.9±0.2	96.9±4.7	98.9±2.1	97.5±4.2
DenseNet [27]	96.2±1.3	98.8±0.4	96.5±1.2	98.8±0.4	96.5±5.4	98.8±2.2	97.3±4.9
AlexNet-STN [23], [30]	92.9±2.1	97.8±0.5	93.4±2.0	97.8±0.5	93.4±7.4	97.8±3.3	94.7±6.9
GoogLeNet-STN [24], [30]	90.7±1.8	97.4±0.3	91.2±1.8	97.4±0.3	91.2±9.8	97.4±3.8	93.1±9.5
VGG-Net-STN [25], [30]	96.8±0.8	99.0±0.3	97.1±0.8	99.0±0.3	97.0±4.4	99.0±1.9	97.7±4.1
ResNet-STN [26], [30]	96.9±0.9	98.9±0.2	97.2±0.9	98.9±0.2	97.2±3.7	98.9±1.8	97.8±3.3
DenseNet-STN [27], [30]	97.0±1.5	99.0±0.4	97.3±1.4	99.0±0.3	97.3±4.4	99.0±1.8	97.9±3.9
G-Net-STN [30]	95.9±1.8	98.7±0.4	96.2±1.6	98.7±0.4	96.2±5.6	98.7±2.2	97.2±5.0
L-Net (Simple)	95.3±0.8	98.3±0.4	95.8±0.7	98.3±0.4	95.8±4.9	98.3±2.5	97.0±4.1
Varifocal-Net (Simple)	96.1±0.6	98.3±0.2	96.5±0.5	98.3±0.2	96.5±4.7	98.3±2.4	96.6±4.7
Varifocal-Net	**98.7±0.7**	**99.2±0.3**	**98.9±0.7**	**99.2±0.3**	**98.9±2.3**	**99.2±1.5**	**99.2±2.1**

To test the usefulness of the varifocal mechanism, we replaced the localization subnet by a simple preprocessing method. The input of the L-Net is not the cropped local region of the original image. Instead, we directly rescaled and padded the minimum bounding box of each chromosome image into the same size (256×256). The processed image contains the whole chromosome part and consequently the extracted features are no longer local. After the L-Net converges, the features learned from the G-Net and the L-Net are concatenated as well for the training of the second stage. We named the L-Net and the Varifocal-Net using such a simple preprocessing step as L-Net (Simple) and Varifocal-Net (Simple), respectively. Table VIII shows that the simple preprocessing method does not facilitate feature learning of fine-grained details. Our method outperforms the L-Net (Simple) and the Varifocal-Net (Simple), which validates the effectiveness of the localization subnet.

Table VIII also provides the results of comparison with other CNN models. To assess our multi-scale feature ensemble strategy, we evaluated the performance of the well-known models that have been proved powerful on the ImageNet dataset, including AlexNet [23], GoogLeNet [24], VGG-Net-D [25], ResNet-101 [26], and DenseNet-121 [27]. The number of convolution layers in these five models and our Varifocal-Net (feature extractor part) are respectively 5, 22, 13, 100, 120, and 28, which are much deeper than previous work in chromosome classification [15], [16]. Besides, we also evaluated Spatial Transformer Network (STN) [30] for performance comparison. It contains 4 Conv layers, 4 Max-Pooling layers, and 2 FC layers. We inserted STN into the first layer of each model and retrained it. Since the parameters of these popular models were compatible with the 3-channel 224×224 natural images (ImageNet), we rescaled our 256×256 grayscale images into 224×224 pixels and then generated 3 channels by directly stacking the original grayscale channel. The preprocessing step was also adopted to normalize all the inputs as mentioned in Sec. III-B. To introduce multi-task learning, we duplicated the classifier settings in each model so that both type and polarity could be predicted at the same time. The loss function is defined as (3) with $\lambda = 0.5$. We trained all models from scratch because the collected samples are sufficient. The results show that all models have acceptable performance. Even the shallowest AlexNet achieved the accuracy (%) of 90.8 and 97.1 for type and polarity classifications, respectively. Among these single-scale CNN models, the highest accuracy and F_1-score were achieved by DenseNet-STN for both type and polarity tasks. However, its result is still inferior to ours, where the error rates of type classification are reduced by half. For the polarity task, our method also outperformed other CNN models. Note that the use of STN does not necessarily improve the performance. Its introduction in GoogLeNet and G-Net brings about obvious decrease.

In the real clinical environment, it is imperative to correctly classify chromosomes having numerical and structural anomalies. To test the robustness of different methods under abnormal circumstance, we specially provide the evaluation results only on unhealthy cases in Table IX. For most CNN-based methods, the performance degraded dramatically on abnormal cases. The AlexNet and GoogLeNet-STN even suffered over 9% loss of accuracy and F_1-score. In contrast, our Varifocal-Net had only a slight performance drop around 1.1% and 0.6% in Acc. per Case of the type and polarity task, respectively. We remarkably outperformed state-of-the-art methods on abnormal chromosome classification.

In Fig. 7, the results of ROC analysis are illustrated for both type and polarity classifications. We first performed ROC analysis per class using a one-vs-all scheme. Then, we averaged all ROC curves over classes and calculated the AUC for each method. It is observed that the proposed Varifocal-Net outperformed other methods with the least false positive predictions and the highest true positive rates. We achieved the highest AUC for both the type and polarity tasks. It demonstrates that in the case of not redesigning a completely brand-new feature extraction architecture, our Varifocal-Net, which

TABLE IX

COMPARISON RESULTS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON UNHEALTHY CASES (MEAN±STANDARD DEVIATION). THE RESULTS ARE PRESENTED IN TERMS OF FOUR EVALUATION METRICS: AVERAGE F_1-SCORE OF ALL TESTING IMAGES (F_1), ACCURACY OF ALL TESTING IMAGES (ACC.), AVERAGE ACCURACY PER PATIENT CASE (ACC. PER CASE), AND AVERAGE ACCURACY PER PATIENT CASE USING THE PROPOSED DISPATCH STRATEGY (ACC. PER CASE-D). (T: TYPE, P: POLARITY.)

Method	F_1 (%)		Acc. (%)		Acc. per Case (%)		Acc. per Case-D (%)
	T	P	T	P	T	P	T
Sharma *et al.* [15]	88.4±3.5	–	88.9±3.5	–	88.9±12.9	–	90.3±12.5
Gupta *et al.* [16]	90.6±1.0	–	90.8±1.0	–	90.8±14.7	–	92.3±14.1
AlexNet [23]	80.7±5.2	93.8±2.2	81.1±5.3	94.0±2.1	81.2±18.6	94.0±7.1	83.6±18.2
GoogLeNet [24]	89.0±4.8	96.1±2.2	89.3±4.7	96.1±2.2	89.2±16.6	96.1±7.5	90.5±16.2
VGG-Net [25]	92.1±2.4	97.6±1.1	92.5±2.5	97.6±1.1	92.5±10.8	97.6±4.0	93.7±10.0
ResNet [26]	93.5±1.7	98.0±1.2	93.8±1.8	98.0±1.2	93.8±9.7	98.0±3.8	94.7±9.0
DenseNet [27]	92.3±2.8	97.7±1.0	92.7±2.7	97.8±1.0	92.6±11.4	97.8±4.0	93.7±10.9
AlexNet-STN [23], [30]	87.2±5.5	95.7±2.1	87.6±5.6	95.8±2.1	87.5±15.1	95.8±6.1	89.2±15.1
GoogLeNet-STN [24], [30]	81.0±4.4	94.2±1.8	81.5±4.6	94.3±1.8	81.5±18.7	94.2±8.6	83.4±19.6
VGG-Net-STN [25], [30]	94.4±2.6	98.4±0.8	94.7±2.5	98.4±0.8	94.7±8.3	98.5±2.9	95.7±8.2
ResNet-STN [26], [30]	96.0±0.5	98.6±0.7	96.2±0.5	98.6±0.6	96.2±5.6	98.6±2.7	96.8±5.0
DenseNet-STN [27], [30]	95.8±2.9	98.5±0.9	96.0±2.8	98.5±0.9	96.0±6.7	98.5±2.4	96.7±6.0
G-Net	95.4±1.4	98.1±0.8	95.6±1.4	98.1±0.7	95.6±7.3	98.1±3.5	96.3±6.6
L-Net	96.6±1.3	98.5±0.5	96.8±1.3	98.6±0.5	96.8±5.8	98.6±2.4	97.5±5.2
G-Net-STN [30]	93.4±3.2	97.8±1.2	93.6±3.0	97.8±1.2	93.6±9.3	97.8±3.5	94.7±8.8
L-Net (Simple)	94.5±1.7	97.7±0.8	94.8±1.6	97.8±0.8	94.8±6.7	97.8±3.0	95.8±6.6
Varifocal-Net (Simple)	95.1±1.1	97.5±0.4	95.3±0.9	97.5±0.4	95.3±5.3	97.5±3.1	95.4±5.3
Varifocal-Net	**97.7±1.6**	**98.6±0.6**	**97.8±1.7**	**98.6±0.6**	**97.8±4.3**	**98.6±2.5**	**98.4±3.9**

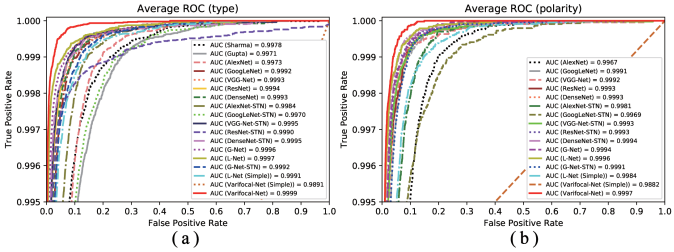


Fig. 7. ROC analysis for the proposed Varifocal-Net and previous CNN models. Each ROC is averaged over all classes and its AUC is calculated. (a) ROC of type classification. (b) ROC of polarity classification.

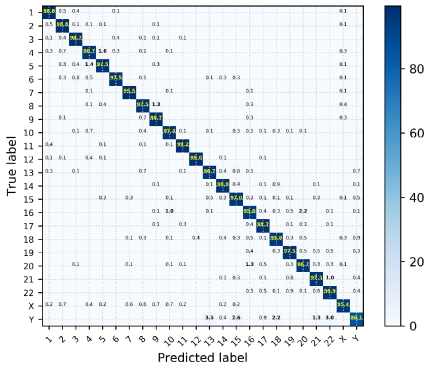


Fig. 8. Confusion matrix of the Varifocal-Net for type classification. The entry in the *i*-th row and *j*-th column denotes the percentage (%) of the testing samples from class *i* that were classified as class *j*. Best viewed magnified.

benefited from the global and local feature ensemble, could further boost the overall classification performance. The lowest three AUCs of the type task were observed for [15], [16], and [23], and simple processing methods, which is consistent with Table VIII. Furthermore, statistical tests were performed using both unpaired and paired t-tests [51], [52]. The Acc. per Case of all five-fold testing samples were tested and the results of two t-tests confirm the significant superiority of the proposed Varifocal-Net against all other methods (p-value $\ll$ 0.05) for both type and polarity tasks.

3) Performance Analysis Results: In this section, we present further experiment results of performance analysis. We computed the confusion matrix to get explanatory insights into the results of type prediction. As shown in Fig. 8, the confusion between class Y and classes No. 13, No. 15, No. 18, No. 21, and No. 22 mainly contributes to the performance drop.

We probed the embedded representations, including the global, local, and concatenated features, in order to illustrate their discrimination capability. We applied the t-SNE [53] approach on testing samples' features to reduce their dimensionality for 2-D visualization. As shown in Fig. 9, the testing samples were clustered by categories and separately dispersed for the concatenated features, with only a small set of samples mixed together. In contrast, for the single-scale global or local features, there exist many large regions where samples of different classes blend together. Compared to Fig. 9(e) and (f), the distance between adjacent clusters in Fig. 9(a)–(d) is smaller. The clusters of global-scale or local-scale features are less compact than that of multi-scale features, making it hard to find a clear boundary for differentiation.

Figs. 10 and 11 illustrate typical examples of correctly and incorrectly classified chromosomes, respectively. Fig. 10 shows that our varifocal mechanism can precisely locate the target region and capture the most discriminative local part with appropriate position and size. For small chromosomes, the predicted box can cover the whole body, while for larger chromosomes, the localization subnet selects partial segments of interest to facilitate accurate recognition. In Fig. 11, misclassified samples are accompanied with their top 5 probabilities for wrong type predictions and 2 probabilities for wrong polarity predictions. It is observed that for most

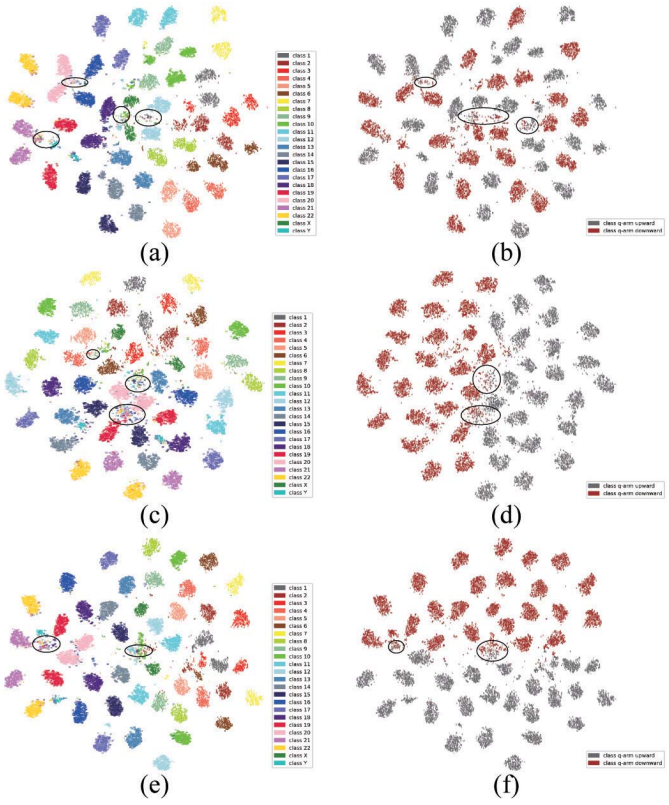


Fig. 9. Feature embedding for chromosomes with t-SNE toolbox [53]. From the perspective of type classification, the global, local, and concatenated features are visualized in (a), (c), and (e), respectively. Similarly, these three features are visualized in (b), (d), and (f) correspondingly for polarity classification. The mixed regions of interest are marked with black circles. Best viewed in color.

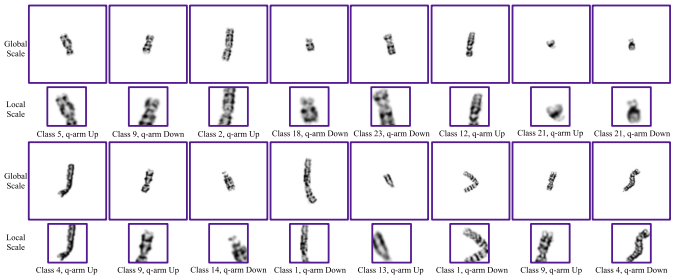


Fig. 10. Examples of correctly classified samples. Both global-scale and local-scale inputs are displayed to visually assess the varifocal mechanism.

incorrect predictions, the probability of the true label ranks just the second highest in order. Besides, some chromosomes are grossly distorted or have unusual shapes of their kinds, increasing the difficulty of accurate classification.

IV. Discussion

In this paper, a three-stage CNN method was proposed for chromosome classification. Its most distinctive characteristics include: 1) the adoption of varifocal mechanism to detect local discriminative regions; 2) the introduction of residual learning and multi-task learning to facilitate feature extraction; 3) the ensemble of global and local features to boost performance;

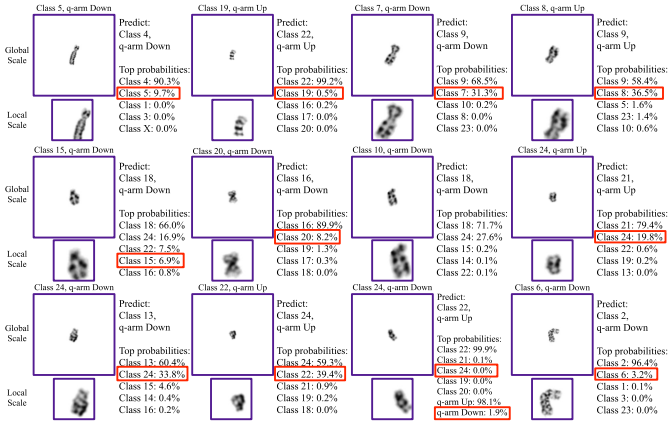


Fig. 11. Examples of misclassified samples. The probabilities of wrong predictions are displayed on the right of each image and each red rectangle encloses the predicted probability of the ground-truth label.

4) the use of a dispatch strategy for type assignment in practical karyotyping per case.

There are mainly two reasons contributing to the inferior performance of the previous CNN-based methods [15], [16]. One is the loss of fidelity caused by the straightening step in their pipelines. Although this step is designed to rectify the shape of chromosomes for normalization, it damages the chromosome's morphological consistency and structural information due to inaccurate medial axis extraction and pixel interpolation. In contrast, the proposed Varifocal-Net is an end-to-end method without any shape correction in advance. The other is the lack of large labeled dataset. Their CNNs, which are designed on small datasets, cannot effectively describe the diversity and variety of chromosomes. Hence, these methods lack generality when evaluated on a large testing set.

As observed from the comparison results in Table VIII, the potent CNN models [23]–[27], [30] performed well because we adapted them into the same settings as ours. In experiments, we adopted multi-task learning and applied necessary normalization on images. Hence, the performance difference among these models, to a certain extent, reflects the difference of their capabilities of global feature extraction. With respect to their accuracy, there exists a bottleneck of improvement for such single-scale models, which inspired us to resort to multi-scale feature ensemble. With the design of the proposed Varifocal-Net, we keep two aims in mind: the excellent feature extraction ability for classification and the strong discrimination of finer regions detected by the localization subnet. Since residual units are employed as the backbone of feature extraction CNNs, the proposed method benefits from the introduction of residual learning. Besides, the multi-task learning strategy also contributes to training the network. For the localization, the varifocal mechanism autonomously focuses on the local part which boosts local feature learning. We can see from Fig. 9 that the integration of both global and local features makes samples of the same category gather closely. It increases between-class distance and reduces chaotic outliers, which explains why our method is superior to the models that only count upon global-scale features.

Additional comparison on unhealthy cases (see Table IX) demonstrated the superior robustness of our method on abnormal chromosome classification. Compared with those single-scale models, our Varifocal-Net utilizes local-scale detail depiction to make up the deficiency of mere consideration of coarse-grained features. Compared to the Varifocal-Net, there does exist a larger performance decrease for the only G-Net and the only L-Net, which confirms the importance of multi-scale feature ensemble strategy. Hence, the proposed method, which possesses excellent generalization abilities, can assist doctors in the clinical karyotyping process where abnormal cases occur from time to time.

The performance improvement of Acc. per Case-D with respect to Acc. per Case in type classification substantiates that the proposed dispatch strategy is effective and suitable for karyotyping within each case. For each method in Table VIII and Table IX, the adoption of the dispatch strategy improves the average accuracy and diminishes the standard deviation for both healthy and unhealthy cases. The generalizability of such strategy lies in the consideration of both maximum likelihood criterion and chromosomal numerical abnormalities.

The proposed method performed less well on chromosomes No. 15, No. 21, and No. 22 (see Table V and Fig. 8) than other classes. Such three kinds of chromosomes are acrocentric and contain a segment called satellite, which is separated from the main body. The shape, size, and orientation of satellites differ from one person to another, thus making it difficult for our model to handle all possible situations. Fig. 8 also shows that chromosome Y is often confused with No. 21 and No. 22. It is because the size and texture of class Y are similar to that of No. 21 and No. 22. Furthermore, the comparatively imbalanced Y samples are not processed with additional data augmentation method, which triggers off poorer recognition of Y. It is noted that although we collected a much larger dataset than previous work, the dataset is still insufficient to cover all possible shapes and sizes of chromosomes. Samples of sex chromosome Y and diversified satellite chromosomes are still in shortage. Therefore for better performance, more data should be collected and generative adversarial networks could be used for sample synthesis in the future.

From the results of Table VII and examples in Fig. 11, it is observed that some long chromosomes (e.g., No. 2 and No. 5) may be misclassified because their long arms tend to bend or distort greatly during the sampling process. Since the proposed method cannot accurately recognize greatly bent chromosomes, future work may involve particular strategies to cope with this situation. Instead of straightening the chromosomes, we might inform the network of the degree of bending deformation by detecting the rotation pivot (e.g., the centromere) and its angle between two arms. Furthermore, for the G-Net and the L-Net, current feature extractor employs the residual block as a backbone. To further improve performance, we may meticulously redesign the network architecture.

V. Conclusion

We have proposed the Varifocal-Net for chromosome classification, which has been evaluated on a large manually constructed dataset. It is a three-stage CNN-based method. The first stage effectively learns global and local features through the G-Net and the L-Net, respectively. Taking a global-scale chromosome image as the input, it precisely detects a local region that is discriminative and abundant in details for further feature extraction. The second stage robustly differentiates chromosomes into various types and polarities via two MLP classifiers. It benefits from multi-scale feature ensemble, with only a few misclassifications. In the third stage, a dispatch strategy was employed to assign each chromosome to a type based on its predicted probabilities. Extensive experimental results demonstrate that our approach outperforms state-of-the-art methods, corroborating its high accuracy and generalizability.

Concerning its role in clinical karyotyping workflow, the Varifocal-Net can accurately perform classification within 1 second after operators manually segment chromosomes of a cell for each patient. The karyotyping result maps it automatically generates offer the possibility for human experts to further check and correct possible misclassifications. Moreover, warnings about possible numerical abnormalities allow operators to pay extra attention to the subsequent diagnosis. The practical use of the Varifocal-Net in the Xiangya Hospital of Central South University suggests its promising potential for alleviating doctors' workload in the diagnosis process.

Acknowledgement

The authors are grateful to the anonymous reviewers for their helpful comments.

References

[1] A. T. Natarajan, "Chromosome aberrations: Past, present and future," *Mutation Res./Fundam. Mol. Mech. Mutagenesis*, vol. 504, nos. 1–2, pp. 3–16, 2002.
[2] A. Theisen and L. G. Shaffer, "Disorders caused by chromosome abnormalities," *Appl. Clin. Genet.*, vol. 3, pp. 159–174, Dec. 2010.
[3] J. Piper, "Automated cytogenetics in the study of mutagenesis and cancer," in *Advances in Mutagenesis Research*. Berlin, Germany: Springer, 1990, pp. 127–153.
[4] E. Schröck *et al.*, "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, no. 5274, pp. 494–497, 1996.
[5] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nature Genet.*, vol. 12, no. 4, p. 368, 1996.
[6] C. Lee *et al.*, "Limitations of chromosome classification by multicolor karyotyping," *Amer. J. Hum. Genet.*, vol. 68, no. 4, pp. 1043–1047, 2001.
[7] D. Huber, L. V. von Voithenberg, and G. V. Kaigala, "Fluorescence *in situ* hybridization (FISH): History, limitations and what to expect from micro-scale FISH?" *Micro Nano Eng.*, vol. 1, pp. 15–24, Nov. 2018.
[8] A. Gozzetti and M. M. Le Beau, "Fluorescence *in situ* hybridization: Uses and limitations," *Seminars Hematol.*, vol. 37, no. 4, pp. 320–333, 2000.
[9] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Medial axis transform-based features and a neural network for human chromosome classification," *Pattern Recognit.*, vol. 28, no. 11, pp. 1673–1683, 1995.
[10] D. Ming and J. Tian, "Automatic pattern extraction and classification for chromosome images," *J. Infr., Millim., THz. Waves*, vol. 31, no. 7, pp. 866–877, 2010.
[11] C. Markou, C. Maramis, A. Delopoulos, C. Daiou, and A. Lambropoulos, "Automatic chromosome classification using support vector machines," in *Pattern Recognition: Methods and Applications*. Hong Kong: iConcept Press, 2012, pp. 1–24.
[12] N. Madian and K. B. Jayanthi, "Analysis of human chromosome classification using centromere position," *Measurement*, vol. 47, pp. 287–295, Jan. 2014.

[13] P. Biyani, X. Wu, and A. Sinha, "Joint classification and pairing of human chromosomes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 2, no. 2, pp. 102–109, Apr. 2005.

[14] F. Abid and L. Hamami, "A survey of neural network based automated systems for human chromosome classification," *Artif. Intell. Rev.*, vol. 49, no. 1, pp. 41–56, 2018.

[15] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, and S. Karande, "Crowdsourcing for chromosome segmentation and deep classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 786–793.

[16] S. Jindal, G. Gupta, M. Yadav, M. Sharma, and L. Vig, "Siamese networks for chromosome classification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2017, pp. 72–81.

[17] Y. Wu, Y. Yue, X. Tan, W. Wang, and T. Lu, "End-to-end chromosome Karyotyping with data augmentation using GAN," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2456–2460.

[18] R. J. Stanley, J. Keller, C. W. Caldwell, and P. Gader, "Centromere attribute integration based chromosome polarity assignment," in *Proc. AMIA Annu. Fall Symp.*, 1996, p. 284.

[19] X. Wang, B. Zheng, S. Li, J. J. Mulvihill, and H. Liu, "A rule-based computer scheme for centromere identification and polarity assignment of metaphase chromosomes," *Comput. Methods Programs Biomed.*, vol. 89, no. 1, pp. 33–42, 2008.

[20] A. S. Arachchige, J. Samarabandu, J. H. M. Knoll, and P. K. Rogan, "Intensity integrated Laplacian-based thickness measurement for detecting human metaphase chromosome centromere location," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 2005–2013, Jul. 2013.

[21] E. Loganathan, M. R. Anuja, and N. Madian, "Analysis of human chromosome images for the identification of centromere position and length," in *Proc. IEEE Point–Care Healthcare Technol. (PHT)*, Jan. 2013, pp. 314–317.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[25] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[27] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jun. 2017, vol. 1, no. 2, p. 3.

[28] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.

[29] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. CVPR*, vol. 2, Jul. 2017, p. 3.

[30] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[31] F. Micci, M. R. Teixeira, and S. Heim, "Complete cytogenetic characterization of the human breast cancer cell line MA11 combining g-banding, comparative genomic hybridization, multicolor fluorescence *in situ* hybridization, RxFISH, and chromosome-specific painting," *Cancer Genet. Cytogenet.*, vol. 131, no. 1, pp. 25–30, 2001.

[32] W. Yang *et al.*, "FISH analysis in addition to G-band karyotyping: Utility in evaluation of myelodysplastic syndromes?" *Leukemia Res.*, vol. 34, no. 4, pp. 420–425, 2010.

[33] E. Rødahl, H. Lybæk, J. Arnes, and G. O. Ness, "Chromosomal imbalances in some benign orbital tumours," *Acta Ophthalmol. Scandinavica*, vol. 83, no. 3, pp. 385–391, 2005.

[34] P. K. Gadhia *et al.*, "A rare double aneuploidy with 48, XXY, +21 karyotype in down syndrome from Gujarat, India," *Int. J. Mol. Med. Sci.*, vol. 4, no. 4, pp. 1–3, 2014.

[35] Y.-S. Fan, V. M. Siu, J. H. Jung, and J. Xu, "Sensitivity of multiple color spectral karyotyping in detecting small interchromosomal rearrangements," *Genet. Test.*, vol. 4, no. 1, pp. 9–14, 2000.

[36] W. J. Godinez, I. Hossain, S. E. Lazic, J. W. Davies, and X. Zhang, "A multi-scale convolutional neural network for phenotyping high-content cellular images," *Bioinformatics*, vol. 33, no. 13, pp. 2010–2019, 2017.

[37] P. Buyssens, A. Elmoataz, and O. Lézoray, "Multiscale convolutional neural networks for vision–based classification of cells," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 342–352.

[38] W. J. Godinez, I. Hossain, and X. Zhang, "Unsupervised phenotypic analysis of cellular images with multi-scale convolutional neural networks," *BioRxiv*, 2018, Art. no. 361410.

[39] X. Pan *et al.*, "Cell detection in pathology and microscopy images with multi-scale fully convolutional neural networks," *World Wide Web*, vol. 21, no. 6, pp. 1721–1743, 2018.

[40] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 588–599.

[41] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 465–469.

[42] W. Lotter, G. Sorensen, and D. Cox, "A multi-scale CNN and curriculum learning strategy for mammogram classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 169–177.

[43] S. Zagoruyko and N. Komodakis. (2016). "Wide residual networks." [Online]. Available: https://arxiv.org/abs/1605.07146

[44] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[45] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[48] J. McGowan-Jordan, A. Simons, and M. Schmid, *International System for Human Cytogenetic Nomenclature*. Basel, Switzerland: Karger, 2016.

[49] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[50] A. Paszke *et al.*, "Automatic differentiation in Pytorch," in *Proc. NIPS-W*, 2017, pp. 1–4.

[51] M. L. Samuels, J. A. Witmer, and A. A. Schaffner, *Statistics for the Life Sciences*, vol. 4. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[52] H. Hsu and P. A. Lachenbruch, "Paired *t* test," *Wiley Encyclopedia of Clinical Trials*. Hoboken, NJ, USA: Wiley, 2007, pp. 1–3.

[53] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.