SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I12P107

Volume 12 Issue 12, 75-87, December 2025

Original Article

Swin-Integrated Cell Detection Model for Embryo Imaging and Morphological Evaluation

R. Barkavi[1], G. Yamuna[2], C. Jayaram[3]

[1,2]*Department of Electronics and Communication Engineering, Annamalai University, Chidambaram, INDIA.*
[3]*Embryologist, Faik IVF Fertility, AL Ain, UNITED ARAB EMIRATES.*

[1]*Corresponding Author : barkavi.radhakrishnan2gmail.com*

Received: 07 October 2025 Revised: 09 November 2025 Accepted: 08 December 2025 Published: 27 December 2025

Abstract *- Accurate analysis of embryonic cell structures is important to evaluate the embryo quality during early developmental stages. The overlapping blastomeres, fragmentation, and irregular morphologies in microscopic images complicate the cell counting and centroid localization procedure. Conventional methods that utilize convolutions and progressive upsampling struggle to model long-range dependencies and adapt to the dense cell regions' spatial irregularities. This research work is proposed with the objective of developing a precise and adaptable model to overcome the limitations in conventional methods. The proposed SwinDePeriNet is a combination of a hierarchical Swin Transformer encoder with a deformable perceptual module, which is specifically developed to capture the dynamic spatial relationships and structural variations. Additionally, a Gaussian-based probabilistic estimator is incorporated to generate localized confidence maps for accurate centroid detection. The proposed model is trained and tested using a benchmark cell image dataset and exhibited 95.3% accuracy, 97.6% R^2 score, 26 MAE, 250 MSE, 93.1% perfect localization rate, 5.2 pixels mean Euclidean error, and a false positive rate of 2.6% which is better than conventional models.*

Keywords *- Microscopic Embryo Imaging, Cell Count Regression, Centroid Detection, Framework, Spatial Confidence Mapping, Transformer-Based Encoding, Deformable Context Fusion, Gaussian Localization.*

1. Introduction

In embryology, accurate cell detection during early-stage development is essential to evaluate the quality and viability of the embryo. Microscopic imaging is generally used for detailed observation of embryo morphology. It is specifically helpful during the cleavage stages, where cellular division patterns reveal important developmental indicators. The need for multi-carrier signal processing in biomedical imaging has increased in recent times for enhancing the contrast, localization precision, and structural clarity.

Specifically, multi-carrier techniques decompose the complex embryonic images into frequency sub-bands, which allows better granular detection of overlapping blastomeres, fragmented nuclei, and irregular cellular arrangements [1]. Thus, applying multi-carrier models in the process of cell detection will enhance the visualization and improve the evaluation by extracting and analyzing the spatial and spectral level features.

The conventional embryo image analysis with multi-carrier systems faces issues like channel estimation and detection. Since the cellular features in microscopy images generally exhibit low reflectance, inconsistent illumination, and texture non-uniformity. Thus, it will distort the detection signal across carriers, and also these variations cause inter-channel interference, which makes centroid localization and cell boundary estimation a highly sensitive process. Moreover, the biological complexity and morphological variability of early-stage embryos make the conventional estimation methods produce high false positive rates, especially in densely packed or irregularly shaped blastomere regions. Hence, accurate cell detection requires improved signal recovery strategies and deformable perception mechanisms, which are specifically developed for biological input analysis.

Several detection algorithms have been developed in recent times to mitigate these limitations. Methods such as frequency-based sub-band filtering, pyramid convolutional networks, and wavelet transform-based deconvolution are used in various research works. Specifically, a hierarchical convolutional backbone exhibits better performance when it is combined with progressive upsampling [2]. However, this approach fails to model long-range interactions between densely clustered cells. Few researchers incorporated Fourier-domain CNNs and Laplacian pyramid regression to decompose images into multiple carrier bands for fine-

grained detection [3-5]. However, these methods suffer from poor generalization due to the use of fixed kernel sampling and static receptive fields.

Few hybrid models have also evolved using deformable convolutions to accommodate local variance, but they lag in performance while attending to the morphologically irregular regions [6, 7]. Transformer-based methods in recent times are computationally intensive and inefficient in dense cellular structures [8-10]. These conventional methods' limitations highlight the need for developing a model that should exhibit better tradeoff between the spatial adaptiveness and detection granularity in cell detection.

The research work is aimed at developing a robust cell counting and localization model incorporating frequency-aware representations with spatial deformability. This ensures global contextual awareness and localized precision in embryo image analysis. The perceptual flexibility in the proposed method, with a transformer-guided encoder-decoder, handles the channel estimation issues and improves the morphological interpretability.

To enable global context encoding and to extract the hierarchical frequency features, the encoder network is used. To handle the morphological irregularities, a deformable perceptual decoder is used so that the attention weights are dynamically adjusted across spatial carriers.

Additionally, a Gaussian-based probabilistic estimator is incorporated to produce spatial confidence maps for precise centroid prediction. The Key Contributions of this research work are presented as follows.

- Proposed a hierarchical Swin Transformer-based encoder to extract frequency-aware spatial features for enhanced context modeling in embryo imaging. Also, a deformable perceptual module is introduced to address irregular cell morphologies and adjust attention dynamically based on structural variance.
- A Gaussian probabilistic estimator is incorporated to generate precise confidence maps for dot-level centroid detection, thereby reducing segmentation overhead. Finally, a lightweight and scalable framework is incorporated to maintain high precision across multiple frequency bands.
- Experimental validation on the benchmark cell image dataset demonstrates the superior performance of the proposed model over conventional models in terms of accuracy and centroid localization.

The remaining discussion in the article is arranged in the following order. Section 2 presents a detailed literature review. Section 3 presents the mathematical model for the proposed work. Section 4 presents the results and discussion. Finally, the conclusion is presented in section 5.

2. Related Works

Recent advancements in biomedical imaging introduced various strategies for automatic cell detection, segmentation, and counting. In this section, recent architectures, instance segmentation models, and density regression networks are xanalyzed to address the challenges in complex cell environments. The segmentation and tracking algorithm designed in [11] for analyzing cell migration utilizes time-lapse microscopy with block-matching 3D filtering for noise suppression. To minimize halo interference, k-means clustering is used, and for boundary extraction, active contour models are incorporated. The experimental analysis exhibits the presented model with better error reduction over existing methods. However, the presented approach is highly sensitive to uneven background and illumination variations.

The localization approach presented in [12] incorporated automated and semiautomated modules to count the immune cells. The challenges in multiple cell morphologies and localization are addressed in this research. The experimental analysis highlights that automated methods over semiautomated models in terms of reliability. However, the presented model accuracy is average and needs improvement in the localization process.

To improve individual cell localization, a two-dimensional directional convolutional neural network is employed in [13]. The presented approach initially assigns a unit vector to each pixel with respect to the cell center. The presented model effectively differentiates the overlapping cells where the cell image has low contrast and blurred boundaries. The experimental analysis presents the improved separation accuracy. However, the generalization of the presented model to unseen tissue types remains limited due to dataset-specific training.

The lightweight encoder–decoder model presented in [14] for nuclei instance segmentation predicts the distance transform and nuclear masks for accurate separation of overlapping nuclei. Using structural distance estimation, the presented model attained enhanced instance boundaries without depending on complex post-processing. Experimental evaluation using benchmark datasets demonstrated better performance over state-of-the-art models. However, the presented model exhibits performance variations while handling the extreme staining differences and inconsistent annotations.

The learning network presented in [15] addresses challenges in segmenting fluorescent spots within microscopy cell images. The presented model incorporates Fourier interpolation preprocessing with an enhanced YOLOv8 detection architecture. The use of an upsampling layer fine-tunes the features and supports boundary-aware segmentation at subpixel precision. Compared to the conventional models, the presented approach achieved better

F1-score with an improvement of up to 8.27% across diverse datasets. However, the model's dependency on high-quality fluorescent imaging and computational load may limit deployment in low-resource settings.

The U-Net model reported in [16] incorporates a self-attention mechanism with a conventional U-Net to enable effective spatial feature extraction for cell counting in both 2D and 3D biological images. Additionally, a modified Batch Normalization approach is introduced to stabilize training on limited datasets. Experimental validation benchmark and own datasets confirm the model's superior performance over conventional techniques. However, the GPU memory limitations of the presented model limit the applicability in a real-time environment.

An interactive dual-network model is presented in [17] for automated cell counting. The presented model integrates a density map regression model with a dynamic ground truth generator for optimal supervision. A hierarchical multi-scale attention module is incorporated to enhance the feature extraction and density estimation. Evaluations of benchmark datasets confirmed the improved accuracy over conventional methods. However, the presented model requires fine-tuning for unseen modalities, and its iterative training increases computational demands compared to single-stage networks.

A deep learning model presented in [18] for cell counting, adapting the U-Net for semantic segmentation and integrating distance transform with watershed algorithms. The presented model effectively handles low-contrast and label-free tissue images. Experimental validation using human brain tumors exhibits better AUC and a correlation coefficient with histological ground truths. However, the dependency on morphological consistency across samples limits the model's generalization to highly variable tissue types.

The ResUnet model presented in [19] employs a fully convolutional U-Net architecture for binary segmentation-based cell localization in fluorescence microscopy. The presented model introduces weighted boundary maps and noise oversampling to improve detection performances. The experimental results demonstrate the model's superior performance with better F1-score and mean absolute error. However, the model's dependency on dataset-specific features may reduce adaptability to varying imaging modalities and staining conditions.

A YOLOv5 model is presented in [20] for automatic cell recognition and counting in Neubauer chamber images. The presented model is pretrained and fine-tuned using 21 annotated lab images. The experimental evaluation exhibits the accuracy and precision of the presented model, which is better than conventional U-Net and openCV approaches. Although the performance is better, the limitation of dataset size and dependency on transfer learning limits the generalizability of models across diverse microscopy image analysis.

2.1. Research Gap

The brief literature review clearly highlights the challenges in cell counting and localization across various methods. The literature analysis confirms that the conventional directional field-based separation and distance transform models are effective. However, these methods struggle to exhibit better performance when the input is dense and has overlapping regions, or when low contrast boundaries exist. Morphological tracking methods and Fourier-interpolated segmentation approaches exhibit better improvements.

However, the scalability of these models to large and heterogeneous datasets is limited. The attention-enhanced architectures U-Net and interactive dual-networks introduce precision enhancements. However, these models demand extensive training data and are highly sensitive to varying cell sizes and densities. In complex images, the cell counting process highlights the need for models that should generalize across different dataset domains. The object detection algorithms lag in performance if the input has combined structures and irregular shapes. These limitations highlight the need for an adaptive model that should handle the spatial deformities, dense clusters, and domain shifts.

3. Proposed Work

The proposed SwinDePeriNet model incorporates multiple models, such as Swin Transformer and a Deformable Perceptual Module, for enhancing the cell detection performance. The Swin Transformer in the proposed model extracts the hierarchical patterns and maintains the computational efficiency through shifted window attention. This makes the model well-suited for handling variable-scale embryonic structures. The Deformable Perceptual Module is incorporated to handle the morphological inconsistencies by dynamically adjusting receptive fields based on contextual variance.

Figure 1 depicts the complete overview of the proposed model, which begins with the preprocessing of microscopy images to enhance contrast and normalize intensity levels across channels. These preprocessed images are passed through the Swin Transformer encoder to obtain multi-level spatial representations. Further, it is passed through the deformable module, which aligns these features and highlights the spatial deformations and fragmented regions. A Gaussian-based probabilistic estimator then processes these features to generate a confidence map, from which centroids are extracted. The final output includes cell count and precise localization, supporting morphological evaluation.

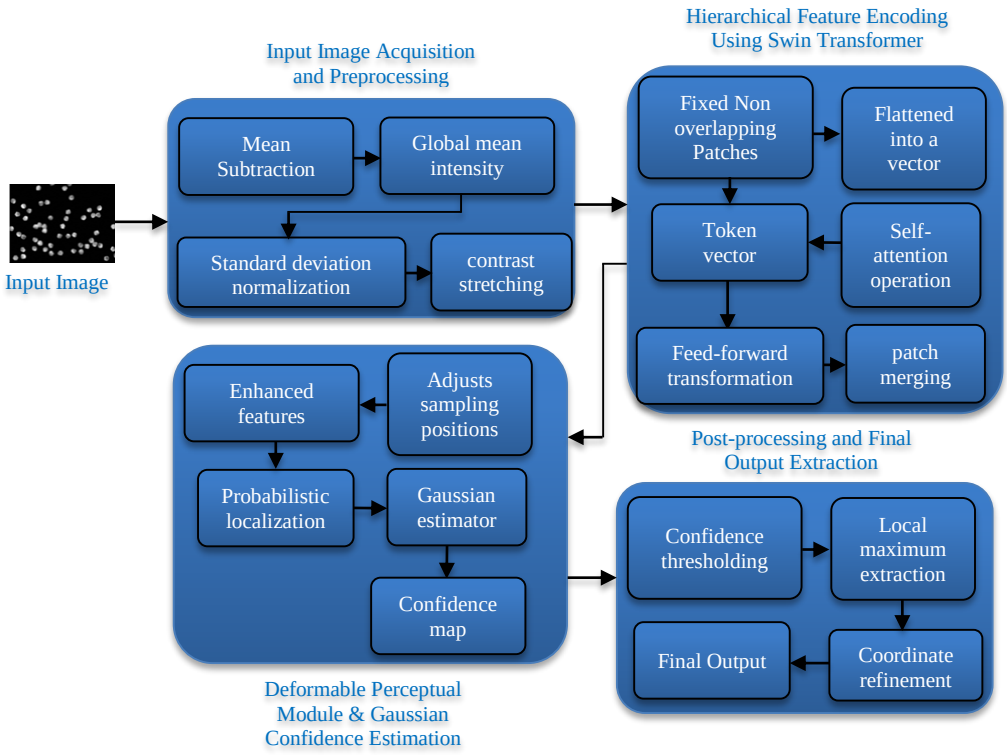

Fig. 1 Proposed model overview

3.1. Input Image Acquisition and Preprocessing

The initial stage of the SwinDePeriNet acquires and preprocesses the embryo microscopy images. Consider the raw microscopy image, denoted by $\mathcal{I} \in .R^{H\times W\times C}$ Where $\mathcal{I}$ is the raw input embryo image captured through a microscope, H and W represent the height and width in pixels, respectively, C indicates the number of image channels. Since the raw image may include intensity disparities and background noise, the first operation is mean subtraction to center the pixel intensity distribution. This is achieved using.

$$\mathcal{I}_c(x,y) = \mathcal{I}(x,y) - \mu_{\mathcal{I}} \tag{1}$$

Where $\mathcal{I}_c(x,y)$ represents the centered pixel intensity, $\mu_{\mathcal{I}}$ is the entire image global mean intensity. Mathematically, it is formulated as

$$\mu_{\mathcal{I}} = \frac{1}{HWC}\sum_{c=1}^{C}\sum_{x=1}^{H}\sum_{y=1}^{W}\mathcal{I}(x,y,c) \tag{2}$$

This operation reduces the bias that is introduced by background illumination or varying image acquisition settings. Followed by standard deviation normalization, which is applied to scale the image intensities to a consistent range. The normalization procedure is mathematically formulated as

$$\mathcal{I}_n(x,y) = \frac{\mathcal{I}_c(x,y)}{\sigma_{\mathcal{I}}} \tag{3}$$

Where $\mathcal{I}_n(x,y)$ is the normalized pixel value, $\sigma_{\mathcal{I}}$ is the standard deviation of image intensities across all pixels and channels, calculated as

$$\sigma_{\mathcal{I}} = \sqrt{\frac{1}{HWC}\sum_{c=1}^{C}\sum_{x=1}^{H}\sum_{y=1}^{W}(\mathcal{I}(x,y,c) - \mu_{\mathcal{I}})^2} \tag{4}$$

This ensures that the pixel values are distributed around zero with unit variance and exhibit stable convergence during learning. To further enhance the visual features for the boundaries and blastomeres textures, contrast stretching is applied on $\mathcal{I}_n$ Using intensity percentile limits. Mathematically, it is expressed as:

$$\mathcal{I}_s(x,y) = \frac{\mathcal{I}_n(x,y) - p_l}{p_h - p_l} \tag{5}$$

Where $\mathcal{I}_s(x,y)$ Is the contrast-stretched output pixel, p_l and p_h are the lower and upper percentile limits of $\mathcal{I}_n$, Values below p_l are clipped to 0, and above p_h are clipped to 1. The final result of this preprocessing stage is the preconditioned image $\mathcal{I}_s \in R^{H\times W\times C}$, which contains uniform illumination, enhanced boundaries, and a standardized dynamic range. This image is then forwarded to the patch partitioning module in the encoder stage for feature extraction.

3.2. Hierarchical Feature Encoding using Swin Transformer

Following preprocessing, the enhanced image $\mathcal{I}_s \in R^{H\times W\times C}$ It is partitioned into fixed-size, non-overlapping patches. Each patch $\mathcal{I}_s^{(i)} \in R^{p\times p\times C}$ is then flattened into a

vector and transformed into an embedding vector through a learned projection.

$$z_0^i = W_p \cdot Flatten\left(I_s^{(i)}\right) + b_p \quad (6)$$

Where $z_0^i \in R^d$ Is the embedding vector for the i^{th} image patch, $W_p \in R^{p^2C \times d}$ Is the learnable weight matrix for patch embedding $b_p \in R^d$ Is the bias vector for embedding, d: Dimension of the token embedding, Flatten$(\cdot)$ It is the operator that converts a patch into a one-dimensional vector. These embedding vectors are grouped into square regions called windows of size $M \times M$. For each window, a self-attention operation is computed to capture the interaction between each pair of patches within that window. For a given query vector $q_i \in R^d$, the attention output is calculated as:

$$o_i = \sum_{j=1}^{M^2} \alpha_{ij} \cdot v_j \quad (7)$$

Where $o_i \in R^d$ is the output vector for query i, α_{ij} Is the attention score between query i and key j, computed as:

$$\alpha_{ij} = \frac{exp\left(\frac{q_i^\top \cdot k_j + b_{ij}}{\sqrt{d}}\right)}{\sum_{l=1}^{M^2} exp\left(\frac{q_i^\top \cdot k_l + b_{il}}{\sqrt{d}}\right)} \quad (8)$$

Where $k_j \in R^d$ Is the key vector for token j, $v_j \in R^d$ Is the value vector for token j, $b_{ij} \in R$ is the learnable relative positional bias between positions i and j, $\sqrt{d}$ Is the scaling factor to stabilize the gradient for each vector q_i, k_j, v_j is obtained from the token embeddings using separate learned linear projections.

$$q_i = W_q \cdot z_i, \quad k_j = W_k \cdot z_j, \quad v_j = W_v \cdot z_j \quad (9)$$

Where $W_q, W_k, W_v \in R^{d \times d}$ It is the learnable transformation matrix. After attention, each token vector is refined by passing it through a feed-forward transformation consisting of two linear mappings with a non-linear activation in between:

$$z_{out}^i = W_2 . \phi(W_1 \cdot o_i + b_i) + b_2 \quad (10)$$

Where $W_1 \in R^{d \times 4d}, W_2 \in R^{4d \times d}$ is the learnable weights for the two fully connected layers, $b_1 \in R^{4d}, b_2 \in R^d$ is the bias vectors, $\phi(\cdot)$ Is the non-linear activation function, typically Gaussian Error Linear Unit (GELU). To extend the receptive field and capture cross-window relationships, a shift in the spatial location of each window is applied in alternate layers. This ensures that each token is allowed to interact with tokens in neighboring windows, without increasing computational complexity. Following the attention and token update, a patch merging operation is conducted at the end of each stage to reduce the spatial resolution and expand the feature depth. For four adjacent patches z_1, z_2, z_3, z_4, their merged representation is given by

$$z_m = W_m \cdot Concat(z_1, z_2, z_3, z_4) \quad (11)$$

Where $z_m \in R^{2d}$ is the merged output vector, $W_m \in R^{4d \times 2d}$ is the linear projection matrix after Concatenation, $Concat(\cdot)$ Indicates the Concatenation of four vectors into one. These encoded features are then forwarded to the deformable perception module for adaptive spatial refinement.

3.3. Deformable Perceptual Module

The encoded features obtained from the Swin-based hierarchical transformer are processed further through a deformable perceptual module. For each output position (x, y) in the deformable feature map, a set of offset vectors $\{(\delta x_k, \delta y_k)\}_{k=1}^K$ It is learned that K represents the number of sampling points in the neighborhood. These offsets adaptively relocate the sampling coordinates relative to the central pixel. The deformable feature at position (x, y) is computed as:

$$\mathcal{F}_d(x, y) = \sum_{k=1}^K w_k(x, y) \cdot F_{enc}(x + \delta x_k, y + \delta y_k) \quad (12)$$

Where $\mathcal{F}_d(x, y)$ Is the output feature at the spatial coordinate (x, y) from the deformable module, F_{enc} Is the input feature map from the encoder, $\delta x_k, \delta y_k$ is the learnable spatial offsets for sampling point k, $w_k(x, y)$ Is the attention-based weight assigned to the k^{th} sampling point at the location (x, y), K is the total number of sampling locations in the neighborhood. These weights w_k They are derived by applying a normalization operation over the local region to ensure the weighted sum is stable and context-sensitive.

$$w_k(x, y) = \frac{\exp(s_k(x,y))}{\sum_{j=1}^K \exp(s_j(x,y))} \quad (13)$$

Where $s_k(x, y)$ Is the raw importance score generated by a learnable function for sampling point k. The offset values $\delta x_k, \delta y_k$ They are not static but predicted from the feature context using a separate convolutional block. Let $\Delta \in R^{H' \times W' \times 2K}$ Represent the offset tensor, where the channel dimension stores both horizontal and vertical displacements for all K points:

$$\Delta(x, y) = W_\Delta * F_{enc}(x, y) \quad (14)$$

Where W_Δ Is the learnable kernel responsible for predicting offsets, $*$ indicating the Convolution operation, $\Delta(x, y)$ indicates the contents $[\delta x_1, \delta y_1, \ldots, \delta x_K, \delta y_K]$ at position (x, y). To handle non-integer positions resulting from offsets, bilinear interpolation is applied when sampling. F_{enc}. If the offset leads to fractional coordinates $(x', y') = (x + \delta x_k, y + \delta y_k)$, then the interpolated value is computed by:

$$F_{enc}(x', y') = \sum_{m=0}^{1} \sum_{n=0}^{1} \gamma_{mn} \cdot F_{enc}(\lfloor x' \rfloor + m, \lfloor y' \rfloor + n) \quad (15)$$

Where γ_{mn} Indicates the bilinear interpolation coefficient. This deformable mechanism allows the network to highlight the regions with curvature and the changes in texture. The result is a refined representation. $\mathcal{F}_d \in R^{H' \times W' \times D}$ That preserves biological fidelity and reduces the spatial noise. Finally, the enhanced features $\mathcal{F}_d$ They are forwarded to the probabilistic localization block, in which a confidence map is generated using a Gaussian estimator to obtain the centroid positions without performing complete segmentation.

3.4. Gaussian Confidence Estimation

In this stage, the encoded feature maps, which have the fine-tuned spatial information, are processed for Gaussian confidence estimation. The proposed model utilizes a probabilistic approach, which generates a continuous value confidence map to highlight each pixel's likelihood. Using a differentiable Gaussian distribution, a dot-level annotation is performed. The confidence map generates a dense, smooth prediction surface, which enables better localization even in blurred or overlapping cell structures. Let the ground truth contain N annotated centroids, each located at spatial coordinates. (x_n, y_n), where $n \in \{1, 2, \ldots, N\}$. The Gaussian confidence value at each pixel position (x, y) In the image, it is formulated as

$$\mathcal{G}(x, y) = \sum_{n=1}^{N} \exp\left(-\frac{(x-x_n)^2 + (y-y_n)^2}{2\sigma^2}\right) \quad (16)$$

Where $\mathcal{G}(x, y)$ indicates the confidence score at the location (x, y), N indicates the total number of annotated centroids in the training image, (x_n, y_n) is the coordinates of the n^{th} cell centroid, σ indicates the standard deviation of the Gaussian kernel, which controls the spatial spread around each centroid. During training, the model learns to regress a predicted confidence map. $\hat{\mathcal{G}}(x, y)$ from the refined feature map $\mathcal{F}_d$. A prediction head composed of a convolutional layer is used to map the deep features into a single-channel output representing the estimated confidence for each pixel. Let this mapping be defined as:

$$\hat{\mathcal{G}}(x, y) = w_g * \mathcal{F}_d(x, y) + b_g \quad (17)$$

Where $\hat{\mathcal{G}}(x, y)$ Is the predicted confidence score at the location (x, y), w_g Are the convolutional kernel weights for Gaussian map prediction, b_g Is the Bias term associated with the convolution, $*$: Denotes the convolution operation. To ensure alignment with ground truth and guide model optimization, a loss function is computed between the predicted map and the ground truth. $\hat{\mathcal{G}}(x, y)$ and the ground truth Gaussian map $\mathcal{G}(x, y)$. The error is measured using the squared Euclidean distance over all pixel positions, given by

$$\mathcal{L}_{map} = \frac{1}{H'W'} \sum_{x=1}^{H'} \sum_{y=1}^{W'} \left(\hat{\mathcal{G}}(x, y) - \mathcal{G}(x, y)\right)^2 \quad (18)$$

Where $\mathcal{L}_{map}$ Is the mean-squared error loss between predicted and accurate confidence maps, H', W' Is the spatial dimension of the output map. After training, during inference, the predicted confidence map $\hat{\mathcal{G}}(x, y)$ It is post-processed to extract the most probable centroid locations. This is achieved by identifying local maxima in the map and applying a confidence threshold τ to suppress spurious peaks. The final set of detected centroids $\mathcal{C}$ is obtained by:

$$\mathcal{C} = \{(x, y) \mid \hat{\mathcal{G}}(x, y) \geq \tau \quad "and \quad " \hat{\mathcal{G}}(x, y) > \hat{\mathcal{G}}(x^{\wedge\prime}, y^{\wedge\prime}) \; \forall \; (x^{\wedge\prime}, y^{\wedge\prime}) \in \mathcal{N}(x, y) \} \quad (19)$$

Where $\mathcal{C}$ is the final predicted set of centroid coordinates, τ is the confidence threshold, $\mathcal{N}(x, y)$ Is the local neighborhood around (x, y) For identifying maxima.

3.5. Loss Formulation and Training Objective

To ensure the SwinDePeriNet model accurately learns to localize cell centroids from embryo microscopy images, a loss function must be defined that penalizes deviations between predicted outputs and known ground truth. Since the model is trained to generate continuous-valued confidence maps rather than categorical labels or masks, a pixel-wise regression loss is more suitable than classification-based metrics. The primary learning objective is to minimize the squared error between the predicted Gaussian response map and the annotated ground truth map constructed using dot-level centroid labels. Let $\hat{\mathcal{G}}(x, y)$ Represent the predicted confidence value at the pixel coordinate. (x, y), and $\mathcal{G}(x, y)$ Denote the corresponding ground truth value obtained using Gaussian kernels centered on annotated centroids. The regression loss for a single image is calculated using the mean of squared differences over all pixels in the output map.

$$\mathcal{L}_{total} = \frac{1}{H' \cdot W'} \sum_{x=1}^{H'} \sum_{y=1}^{W'} \left(\hat{\mathcal{G}}(x, y) - \mathcal{G}(x, y)\right)^2 \quad (20)$$

Where $\mathcal{L}_{total}$ Is the total loss used to train the model, (H', W') Is the height and width of the predicted confidence map, $\hat{\mathcal{G}}(x, y)$ Is the predicted confidence score at the pixel (x, y), $\mathcal{G}(x, y)$ Is the ground truth Gaussian value at the pixel (x, y). This formulation encourages the model to output values close to one near annotated centroids and values approaching zero elsewhere. Since the Gaussian peaks are spatially continuous, the model also learns to approximate the shape and spread of accurate centroids, enabling high-resolution localization without producing segmentation boundaries. In some scenarios, especially when cells are densely packed, background pixels may significantly outnumber foreground peaks, leading to class imbalance in the regression space. To handle this, a spatial weighting mask $\mathcal{W}(x, y)$ It is incorporated, which highlights the high-

confidence zones during learning. The revised loss function is then formulated as follows.

$$\mathcal{L}_{weighted} = \frac{1}{H' \cdot W'} \sum_{x=1}^{H'} \sum_{y=1}^{W'} \mathcal{W}(x,y) \cdot \left(\hat{\mathcal{G}}(x,y) - \mathcal{G}(x,y)\right)^2 \quad (21)$$

Where $\mathcal{W}(x,y)$ Represents the weighting function. The final objective is to minimize $\mathcal{L}_{total}$ Over the entire training set. For a dataset containing N_I Images, the total batch loss is formulated as:

$$\mathcal{L}_{batch} = \frac{1}{N_I} \sum_{i=1}^{N_I} \mathcal{L}_{total}^{(i)} \quad (22)$$

Where $\mathcal{L}_{batch}$ represents the final averaged loss, N_I represents the current batch image count, $\mathcal{L}_{total}^{(i)}$ represents the loss computed for the i^{th} training image.

3.6. Post-Processing and Final Output Extraction

After the confidence map $\hat{\mathcal{G}}(x,y) \in R^{H' \times W'}$ The next process is to convert the continuous-valued heatmap into a discrete set of centroid coordinates. In the post-processing, a predefined confidence threshold ($\tau \in [0,1]$) It is used to eliminate low-confidence regions. The thresholded binary mask $\mathcal{B}(x,y)$ is formulated as

$$\mathcal{B}(x,y) = \begin{cases} 1 & if\ \hat{\mathcal{G}}(x,y) \geq \tau \\ 0 & otherwise \end{cases} \quad (23)$$

Where $\mathcal{B}(x,y)$ indicates the binary indicator for the candidate centroid at pixel (x,y), τ is the minimum confidence, $\hat{\mathcal{G}}(x,y)$ Represents the predicted confidence score.

Further, the local maxima are identified to ensure that prominent peaks in each region are selected. Let $\mathcal{N}(x,y)$ Denote the square neighborhood of size. $(r \times r)$ pixels centered at (x,y). A location is considered a local maximum if its value is greater than or equal to all others in the defined window.

$$(x,y) \in \mathcal{C} \quad \text{if} \quad \hat{\mathcal{G}}(x,y) = \max_{(u,v) \in \mathcal{N}(x,y)} \hat{\mathcal{G}}(u,v) \quad \text{and} \quad \mathcal{B}(x,y) = 1 \quad (24)$$

Where $\mathcal{C}$ represents the final set of predicted centroid coordinates, $\mathcal{N}(x,y)$ Indicates the local neighborhood around pixel, r indicates the square window side length. This ensures that the final output has only confident and isolated peaks. Further, a coordinate adjustment step is applied to fine-tune the predictions. Let $\mathcal{R}_k$ represents a region around the k^{th} detected maximum (x_k, y_k), then the refined centroid $(\tilde{x}_k, \tilde{y}_k)$ is computed as

$$\tilde{x}_k = \frac{\sum_{(u,v) \in \mathcal{R}_k} u \cdot \hat{\mathcal{G}}(u,v)}{\sum_{(u,v) \in \mathcal{R}_k} \hat{\mathcal{G}}(u,v)}, \tilde{y}_k = \frac{\sum_{(u,v) \in \mathcal{R}_k} v \cdot \hat{\mathcal{G}}(u,v)}{\sum_{(u,v) \in \mathcal{R}_k} \hat{\mathcal{G}}(u,v)} \quad (25)$$

Where $\tilde{x}_k, \tilde{y}_k$ represents the refined coordinates, $\mathcal{R}_k$ Represents the local region around the detected peak for interpolation, u, v indicates the pixel coordinates, $\hat{\mathcal{G}}(u,v)$ The confidence values are used as weights. This interpolation-based fine-tuning reduces the quantization errors introduced in earlier stages of the process. Additionally, it enhances sub-pixel localization accuracy, which is crucial in microscopy image analysis. The final output of this step is the complete list of predicted centroids, which is mathematically expressed as:

$$\mathcal{C}_{final} = \left\{(\tilde{x}_k, \tilde{y}_k); |; k = 1,2, \ldots, N_p\right\} \quad (26)$$

Where $\mathcal{C}_{final}$ represents the set of all predicted cell centroids, N_p Represents the total number of final detections. The summarized pseudocode for the proposed model is presented as follows.

Pseudocode: SwinDePeriNet – Cell Centroid Detection

Input: ($\mathcal{I} \in R^{H \times W \times C}$) — Raw embryo image

Output: $\mathcal{C}_{final} = (\widetilde{x_1}, \widetilde{y_1}), (\widetilde{x_2}, \widetilde{y_2}), \ldots, (\widetilde{x_N}, \widetilde{y_N})$ — Detected centroid coordinates

Initialization: Set patch size p, window size M, embedding dimension d, define Gaussian spread σ, Set threshold ($\tau \in [0,1]$) for post-processing, Initialize learnable parameters $(W_p, b_p, W_q, W_k, W_v, W_\Delta, W_g)$

Begin

Compute global mean $\mu \leftarrow \text{Mean}(\mathcal{I})$

Compute standard deviation $\sigma_{\mathcal{I}} \leftarrow \text{StdDev}(\mathcal{I})$

Normalize image $\mathcal{I} * n(x,y) = \frac{\mathcal{I}(x,y) - \mu}{\sigma * \mathcal{I}}$

Apply contrast stretching

Divide $\mathcal{I}_n$ into non-overlapping patches of size $(p \times p)$

For each patch i, compute the embedding. $z_0^i = W_p \cdot Flatten\left(\mathcal{I}_n^{(i)}\right) + b_p$

Organize tokens into windows of size. $(M \times M)$

For each window

Compute self-attention weights using projected queries, keys, and values.

Apply relative positional encoding.

Aggregate window responses

Apply shifted windows in alternate layers.

Perform patch merging to reduce spatial size and increase depth.

Output encoded feature map $\mathcal{F}_{enc}$

Predict spatial offsets $\Delta(x,y) = W_\Delta * \mathcal{F}_{enc}(x,y)$

For each pixel (x,y), initialize $\mathcal{F}_d(x,y) \leftarrow 0$

Loop over $k = 1$ to K

Compute the deformed location. $(x_k, y_k) = (x + \delta x_k, y + \delta y_k)$

Interpolate $\mathcal{F}_{enc}(x_k, y_k)$ using the bilinear method

Weight using learned scalar w_k

Accumulate $\mathcal{F}_d(x,y) += w_k(x,y) \cdot \mathcal{F}_{enc}(x_k, y_k)$

Convolve $(\mathcal{F}_d)$ to produce a prediction $\hat{\mathcal{G}}(x,y) = W_g * \mathcal{F}_d(x,y) + b_g$
Construct ground truth Gaussian map $\mathcal{G}(x,y)$ using annotated centroids
Compute loss $\mathcal{L} = \frac{1}{H'W'}\sum_{x,y}\left(\hat{\mathcal{G}}(x,y) - \mathcal{G}(x,y)\right)^2$
Initialize set $\mathcal{C}_{final} \leftarrow \emptyset$
For all (x,y)
If $\hat{\mathcal{G}}(x,y) \geq \tau$
Check if $\hat{\mathcal{G}}(x,y)$ is maximum in the neighborhood
If yes, define local region $\mathcal{R}$
Compute refined position
$\tilde{x} = \frac{\sum_{(u,v)\in\mathcal{R}} u\cdot\hat{\mathcal{G}}(u,v)}{\sum_{(u,v)\in\mathcal{R}}\hat{\mathcal{G}}(u,v)}$ $\tilde{y} = \frac{\sum_{(u,v)\in\mathcal{R}} v\cdot\hat{\mathcal{G}}(u,v)}{\sum_{(u,v)\in\mathcal{R}}\hat{\mathcal{G}}(u,v)}$
Append $(\tilde{x},\tilde{y})$ to $\mathcal{C}_{final}$
Return
End

4. Results and Discussion

The proposed SwinDePeriNet's experimentation is validated through a Python tool that incorporates CUDA-enabled GPU acceleration for improved training and testing. The benchmark dataset used in the proposed model includes cell images from the Kaggle repository. The entire dataset is divided into a 70% training set and a 30% testing set. The preprocessing steps include normalization, contrast enhancement, and resolution standardization. The training used Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8. For better validation, the proposed model is compared with existing methods, and the details of simulation hyperparameters are presented in Table 1 for all the models.

The proposed model's experimentation is validated using a benchmark cell image dataset available in the Kaggle repository [21]. The dataset has synthetically generated microscopy images that replicate the visual and structural characteristics of biological cell cultures. The images in the dataset are also accompanied by their binary masks, which clearly display the individual cell boundaries. All the input images are preprocessed and processed through the proposed modules. A summary of the dataset is presented in Table 2.

Table 1. Simulation hyperparameters

S.No	Method	Parameter	Value
1	Proposed SwinDePeriNet	Learning Rate	0.0001
2		Batch Size	8
3		Optimizer	Adam
4		Patch Size	4 × 4
5		Window Size	7 × 7
6		Depth of Transformer	4
7		Number of Attention Heads	6
8		Epochs	200
9		Weight Decay	1e-4
10		Loss Function	Density + Content
11	Cell-Net	Learning Rate	0.001
12		Batch Size	8
13		Optimizer	Adam
14		Epochs	200
15		Atrous Dilation Rates	[1, 2, 3]
16		Pyramid Levels	3
17		Residual Blocks	5
18	CSRNet	Learning Rate	0.0005
19		Batch Size	8
20		Optimizer	SGD
21		Momentum	0.9
22		Epochs	200
23		Backbone	VGG-16
24	MCNN	Learning Rate	0.0001
25		Batch Size	8
26		Optimizer	Adam
27		Epochs	200
28		Number of Columns	3
29		Filter Sizes	9, 7, 5

Table 2. Dataset description

Class / Subset	Total Samples	Training Samples (80 %)	Testing Samples (20 %)
All images (cell-count labels)	19,200	15,360	3,840
Images with segmentation masks (foreground/background)	1,200	960	240
Total	20400	16320	4080

The proposed model training and testing performance is presented in Table 3 for the metrics like MAE, MSE, accuracy, R^2 Score, perfect localization rate, and FPR metrics. The results clearly demonstrate the superior performance of the proposed model, with a lower MAE of 6.21 and MSE of 50.3 during training. During testing, the MAE increases to 6.90, and the MSE increases to 59.2.

This indicates the minimal overfitting of the proposed model. The overall accuracy of the proposed model is 96.8% during training and 95.2% during testing. The training and testing values of the perfect localization rate are 94.2% and 93.1%, respectively, whereas the mean Euclidean distance increases from 3.5 to 3.9 pixels. This enhanced performance highlights the better precision of the proposed model.

Table 3. Proposed model training and testing performances

Metric	Training Phase	Testing Phase
Mean Absolute Error (MAE)	6.21	6.90
Mean Squared Error (MSE)	50.3	59.2
Accuracy (%)	96.8	95.2
R^2 Score	0.982	0.976
Mean Euclidean Distance (pixels)	3.5	3.9
Perfect Localization Rate (%)	94.2	93.1
False Positive Rate (%)	2.4	2.6

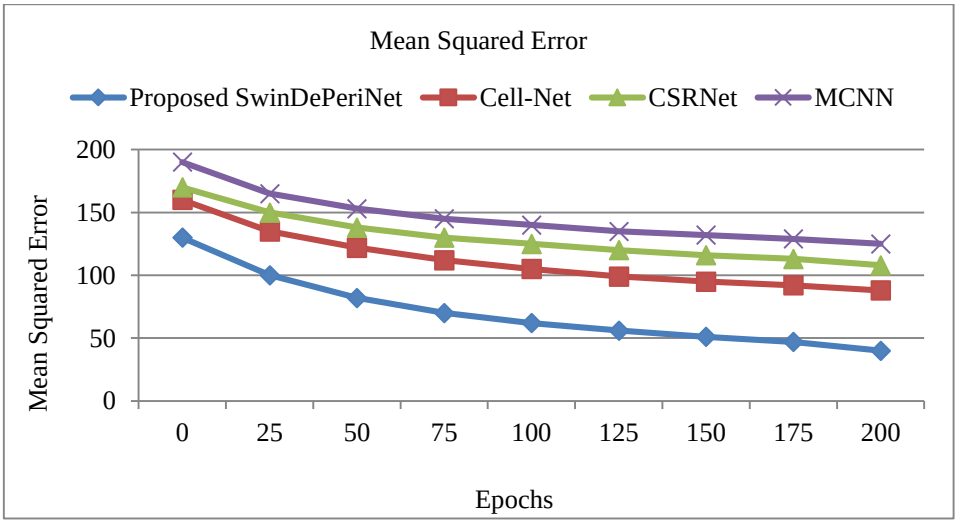

Fig. 2 MSE analysis

To validate the proposed model with existing methods, a detailed comparative analysis is performed. The proposed model's performance is analyzed in comparison to existing methods, including Cell-Net, CSRNet, and MCNN models. The comparative analysis of the mean squared error metric presented in Figure 2 highlights the lower MSE of the proposed SwinDePeriNet model, which is 40.3, compared to the MSE of 91.4%, 109.6%, and 124.1% for the Cell-Net, CSRNet, and MCNN models, respectively. The perfect localization rate analysis presented in Figure 4 highlights the better location ability of the proposed model over conventional methods. The proposed model exhibits a better localization rate of 93.1% for the 200th epoch. In contrast, the existing method, Cell-Net, exhibits 82.1%, CSRNet exhibits 77.1% and MCNN exhibits 74.0% which is lower than the proposed model. The consistent improvement of the proposed model is due to the deformable perceptual encoding and Swin attention modules, which effectively capture the changes and morphological changes in embryo structures.

The R^2 Score comparative analysis across different window sizes is presented in Figure 4 to demonstrate the robustness of the models. The proposed model exhibits an R^2 Score of 0.9756 for the optimal window size of 7. Whereas the existing methods like Cell-Net exhibit 0.9455, CSRNet exhibits 0.9282, and MCNN exhibits 0.9056, which is less than the proposed model. For the other window sizes, like 5, 6, 8, and 9, the proposed model exhibits superior performance compared to existing methods. The consistent performance of the proposed model highlights the model's ability to capture the long-range spatial dependencies in cell structure analysis.

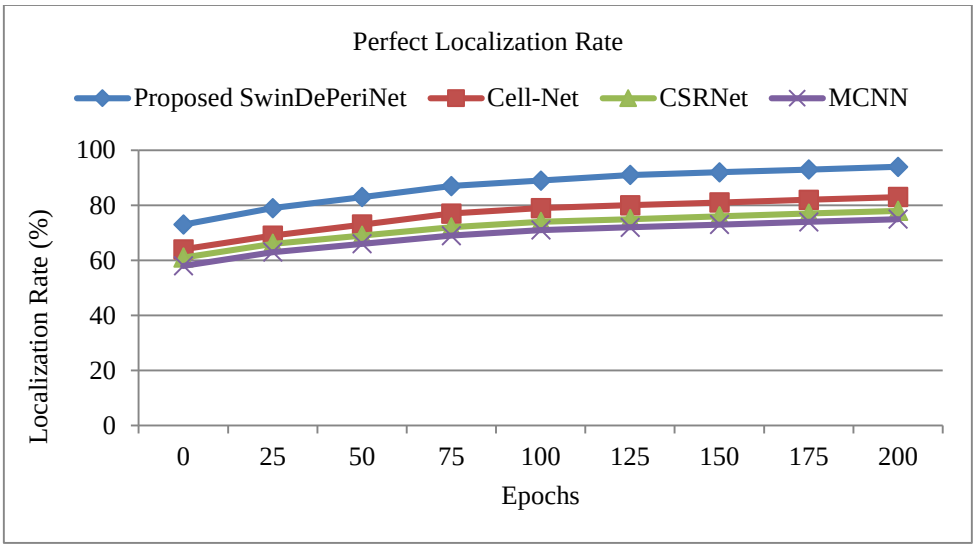

Fig. 3 Perfect localization rate analysis

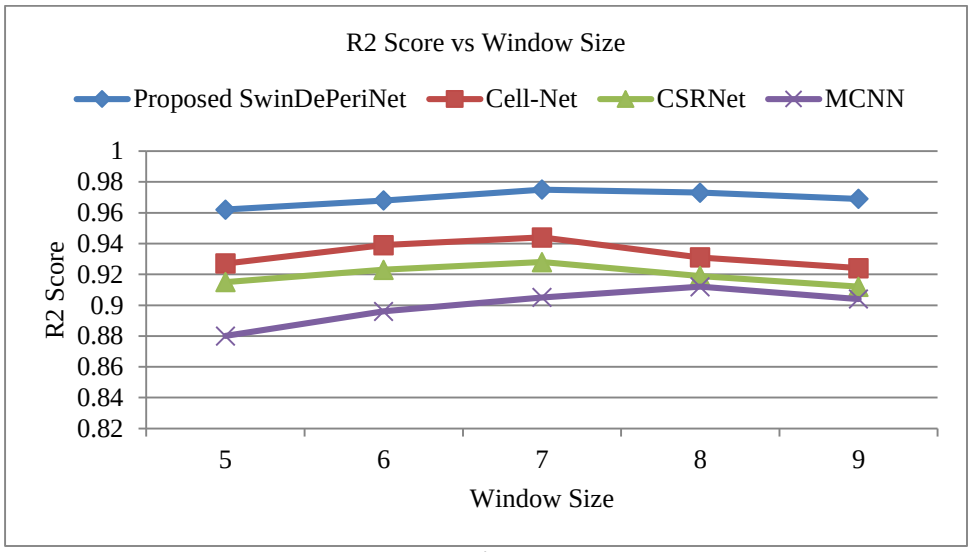

Fig. 4 R^2 Score analysis

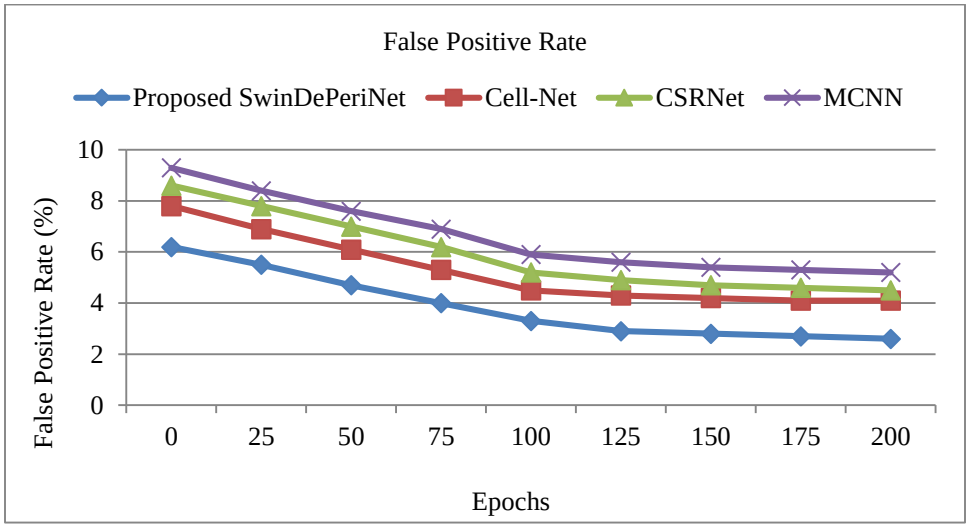

Fig. 5 FPR Analysis

The False Positive Rate (FPR) comparative analysis of the proposed and existing methods is presented in Figure 5. The results clearly present the proposed model's initial lower rate of 6.1% which is better compared to Cell-Net, which exhibits 7.8%, CSRNet, which exhibits 8.5% and MCNN, which exhibits 9.3%. For the maximum epoch, the proposed model reaches an FPR of 2.6% which is better compared to the existing Cell-Net, CSRNet, and MCNN methods' FPR of 4.1%, 4.5% and 5.2% respectively. The better FPR of the proposed model ensures enhanced centroid identification and accurate differentiation between the cell regions.

The variation in localization accuracy is comparatively presented in Figure 6 over patch size. The proposed SwinDePeriNet exhibits better localization precision by exhibiting the lowest Mean Euclidean Distance (MED) of 3.88px. The existing methods like Cell-net, CSRNet, and MCNN exhibit high values of 4.82 px, 5.01 px, and 5.30 px for a patch size of 4. When the patch size is increased to 10, the MCNN reaches a maximum MED of 6.4px, whereas the proposed model maintains a stable precision with 5.22px. These precision results highlight the robustness of the proposed model over spatial resolution changes and enhanced localization accuracy.

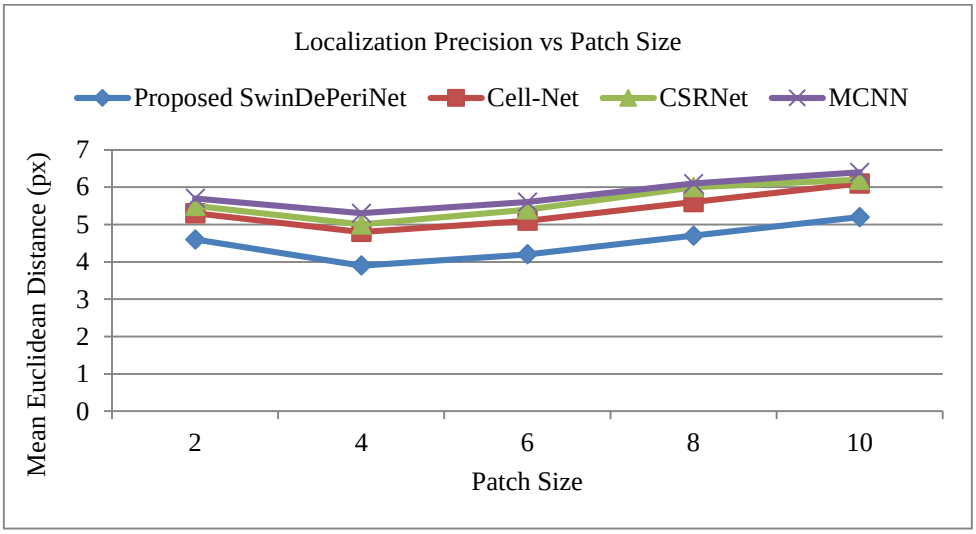

Fig. 6 Localization precision analysis

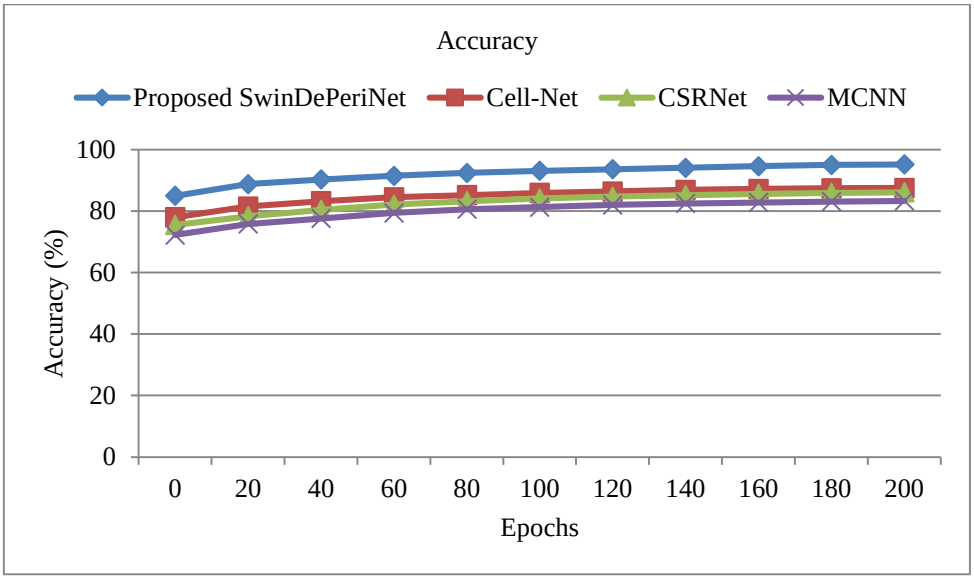

Fig. 7 Accuracy analysis

Figure 7 presents the accuracy comparative analysis of the proposed SwinDePeriNet. The comparative analysis clearly highlights better performance of the proposed model with an accuracy of 95.3% for the maximum epoch. The existing model Cell-Net exhibits an accuracy of 87.1% whereas CSRNet exhibits an accuracy of 86.2% and MCNN exhibits an accuracy of 83.2% which is lower than the proposed model. The better improvement in accuracy of the

proposed model confirms its better convergence behavior and feature extraction ability.

The overall performance analysis presented in Table 4 highlights the proposed model's superior performance for all the metrics. The proposed model exhibits the highest accuracy of 95.3% which is better than cell accuracy of 87.1%, CSR-Net accuracy of 86.2% and MCNN model accuracy of 83.2%. In case of perfect localization rate, the proposed model is superior with 93.1% whereas the existing methods attain in the range of 745 to 82% which is lesser. The regression precision metrics demonstrate the proposed model's better performance by attaining an R^2 score of 97.6% which is better than other existing methods. The performance of the proposed model for the false positive rate is better, with the lowest rate of 2.6% whereas existing methods exhibit a range of 4.1 to 5.2%. Overall, the proposed model's performance metrics are much better than existing methods in cell counting and centroid localization of microscopy images.

Table 4. Overall performance analysis

Metric	Cell-Net	CSRNet	MCNN	Proposed SwinDePeriNet
Accuracy (%)	87.1	86.2	83.2	95.3
Perfect Localization Rate (%)	82.2	77.0	74.0	93.1
R^2 Score (Window Size = 7)	94.5	92.8	90.8	97.6
False Positive Rate (%)	4.1	4.5	5.2	2.6
Mean Euclidean Distance (px)	6.1	6.2	6.4	5.2
Mean Absolute Error (MAE)	35	42	48	26
Mean Squared Error (MSE)	300	350	420	250

5. Conclusion

This research work presents a transformer model, SwinDePeriNet, to count the cells and localize the centroid in microscopy images. The proposed work incorporates the Swin transformer encoder along with a deformable perceptual layer for adaptive feature processing.

The experimentation utilizes the Adam optimizer to train and test the proposed model. Benchmark dataset exhibits the proposed model performance as 95.3% accuracy, 93.1% perfect localization rate, and a R^2 score of 97.6%. Error values were minimized with an MAE of 26, MSE of 250, and Euclidean distance of 5.2 pixels. Compared to existing methods such as Cell-Net, CSRNet, and MCNN, the performance of the proposed model is much better for all the metrics. Apart from the advantages, the proposed model has a minor limitation, as the experimentation utilizes synthetic data. Future work will overcome this limitation by incorporating real cell images and optimization procedures to improve the computational efficiency.

References

[1] Anna Cecchele et al., "Cellular and Molecular Nature of Fragmentation of Human Embryos," *International Journal of Molecular Sciences*, vol. 23, no. 3, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Reza Moradi Rad et al., "Cell-Net: Embryonic Cell Counting and Centroid Localization via Residual Incremental Atrous Pyramid and Progressive Upsampling Convolution," *IEEE Access*, vol. 7, pp. 81945-81955, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Chang Qiao et al., "Evaluation and Development of Deep Neural Networks for Image Super-Resolution in Optical Microscopy," *Nature Methods*, vol. 18, pp. 194-202, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Pavan Chandra Konda et al., "Fourier Ptychography: Current Applications and Future Promises," *Optics Express*, vol. 28, no. 7, pp. 9603-9630, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Nguyen Duy Tan et al., "From Images to Insights: Cell Counting and Uniformity Grading of Day 3 Embryos," *Computers in Biology and Medicine*, vol. 196, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Qingnan Ma et al., "DCO-Net: Deformable Convolution-based O-Shape Network for Fully Automated Placenta Segmentation," *Digital Signal Processing*, vol. 145, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Chengyang Zhang et al., "Difference-Deformable Convolution with Pseudo Scale Instance Map for Cell Localization," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 355-366, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jiazheng Liu et al., "A Robust Transformer-Based Pipeline of 3D Cell Alignment, Denoise and Instance Segmentation on Electron Microscopy Sequence Images," *Journal of Plant Physiology*, vol. 297, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Xi Hu et al., "Transformer-based Deep Learning for Accurate Detection of Multiple Base Modifications using Single Molecule Real-Time Sequencing," *Communications Biology*, vol. 8, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Wided S. Miled et al., "Semantic Segmentation of Human Blastocyst Images using Deep CNNs and Vision Transformers," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Hongju Jo et al., "A Novel Method for Effective Cell Segmentation and Tracking in Phase Contrast Microscopic Images," *Sensors*, vol. 21, no. 10, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Hasita V. Nalluri et al., "Optimizing Colocalized Cell Counting using Automated and Semiautomated Methods," *Neuroinformatics*, vol. 23, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yajie Chen et al., "Cell Localization and Counting using Direction Field Map," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 359-368, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Syed Nauyan Rashid, and Muhammad Moazam Fraz, "Nuclei Probability and Centroid Map Network for Nuclei Instance Segmentation in Histology Images," *Neural Computing and Applications*, vol. 35, pp. 15447-15460, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Huan Liu et al., "UPBAS-Net: An Upsampling-Powered Boundary-Aware Segmentation Network for Fluorescent Spots in Microscopy Images," *Analytical Chemistry*, vol. 97, no. 40, pp. 22200-22210, 2025. [Google Scholar] [Publisher Link]

[16] Yue Guo et al., "SAU-Net: A Unified Network for Cell Counting in 2D and 3D Microscopy Images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 1920-1932, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Rui Liu et al., "Interactive Dual Network with Adaptive Density Map for Automatic Cell Counting," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 4, pp. 6731-6743, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Qianqian Zhang et al., "Automatic Cell Counting from Stimulated Raman Imaging using Deep Learning," *Plos One*, vol. 16, no. 7, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Roberto Morelli et al., "Automating Cell Counting in Fluorescent Microscopy through Deep Learning with C-ResUnet," *Scientific Reports*, vol. 11, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] López Flórez et al., "Automatic Cell Counting with YOLOv5: A Fluorescence Microscopy Approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 3, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Synthetic Cell Images and Masks, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/vbookshelf/synthetic-cell-images-and-masks-bbbc005-v1