frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fnins.2024.1363288

Check for updates

OPEN ACCESS

EDITED BY
Benjamin Thompson,
University of Waterloo, Canada

REVIEWED BY
Da Chen,
Qilu University of Technology, China
Mengting Liu,
Sun Yat-sen University, China
Fan Huang,
The University of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Qinxiang Zheng
✉ zhengqinxiang@aliyun.com
Jiong Zhang
✉ jiong.zhang@ieee.org

RECEIVED 30 December 2023
ACCEPTED 13 March 2024
PUBLISHED 26 March 2024

CITATION
Zhang D, Zhang J, Li S, Dong Z, Zheng Q and Zhang J (2024) U-NTCA: nnUNet and nested transformer with channel attention for corneal cell segmentation.
Front. Neurosci. 18:1363288.
doi: 10.3389/fnins.2024.1363288

U-NTCA: nnUNet and nested transformer with channel attention for corneal cell segmentation

Dan Zhang[1], Jing Zhang[2], Saiqing Li[3,4], Zhixin Dong[3,4], Qinxiang Zheng[3,4,5]* and Jiong Zhang[2,5]*

[1]School of Cyber Science and Engineering, Ningbo University of Technology, Ningbo, China, [2]Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China, [3]National Clinical Research Center for Ocular Diseases, Wenzhou Medical University, Wenzhou, China, [4]The Eye Hospital and School of Ophthalmology and Optometry, Wenzhou Medical University, Wenzhou, China, [5]The Ningbo Eye Hospital of Wenzhou Medical University, Ningbo, China

Background: Automatic segmentation of corneal stromal cells can assist ophthalmologists to detect abnormal morphology in confocal microscopy images, thereby assessing the virus infection or conical mutation of corneas, and avoiding irreversible pathological damage. However, the corneal stromal cells often suffer from uneven illumination and disordered vascular occlusion, resulting in inaccurate segmentation.

Methods: In response to these challenges, this study proposes a novel approach: a nnUNet and nested Transformer-based network integrated with dual high-order channel attention, named U-NTCA. Unlike nnUNet, this architecture allows for the recursive transmission of crucial contextual features and direct interaction of features across layers to improve the accuracy of cell recognition in low-quality regions. The proposed methodology involves multiple steps. Firstly, three underlying features with the same channel number are sent into an attention channel named $g^n Conv$ to facilitate higher-order interaction of local context. Secondly, we leverage different layers in U-Net to integrate Transformer nested with $g^n Conv$, and concatenate multiple Transformers to transmit multi-scale features in a bottom-up manner. We encode the downsampling features, corresponding upsampling features, and low-level feature information transmitted from lower layers to model potential correlations between features of varying sizes and resolutions. These multi-scale features play a pivotal role in refining the position information and morphological details of the current layer through recursive transmission.

Results: Experimental results on a clinical dataset including 136 images show that the proposed method achieves competitive performance with a Dice score of 82.72% and an AUC (Area Under Curve) of 90.92%, which are higher than the performance of nnUNet.

Conclusion: The experimental results indicate that our model provides a cost-effective and high-precision segmentation solution for corneal stromal cells, particularly in challenging image scenarios.

KEYWORDS

cornea, cell segmentation, nested transformer, nnUNet, multi-scale

1 Introduction

Corneal stroma layer comprises collagen fibers, accounting for 90% of the overall thickness of cornea. The corneal stroma cells, as the major cell type of the stroma, produce proteins that provide structure to the stroma and maintain the homeostasis of cornea (Barrientez et al., 2019). The injury of stromal cells tend to cause corneal irreversible damage (Barrientez et al., 2019). Previous studies have shown that the segmentation of corneal stromal cells provide the possibility to quantify cell density and other morphological changes (Arıcı et al., 2014). This process assists ophthalmologists in intuitively acquiring geometric variations to support clinical analysis (Al-Fahdawi et al., 2018). Consequently, it enables the identification of deformities or erosion caused by viruses, helping prevent irreversible pathological damage that could lead to significant visual impairment or even blindness in patients (Subramaniam et al., 2021). In particular, when compared with healthy corneas, keratoconus presents a conical protrusion and the stroma becomes significantly thinner (Lagali, 2020). Thus, the segmentation of stromal cells and the subsequent morphological measurements are helpful for ophthalmologists to judge the severity and progress of the disease.

The utility of automatic cell segmentation approaches significantly enhances the efficiency of ophthalmologists, thereby reducing the dependency on highly experienced experts (Shang et al., 2022). Various widely employed algorithms, including K-means clustering (Yan et al., 2012), edge detection (Pan et al., 2015), and watershed (Sharif et al., 2012) have been utilized to achieve automatic cell segmentation. Among them, watershed stands out due to its ability to identify challenging regions by incorporating distance transform, variance filtering, and gradient analysis (Lux and Matula, 2020). Dagher and El Tom (2008) proposed a hybrid snake-shape parameter optimization by combining the watershed algorithm with active contour, employing region merging and multi-scale techniques to alleviate issues associated with insufficient segmentation. Al-Fahdawi et al. (2018) employed Fourier transform to mitigate image noise and combined watershed for endothelial cell boundary detection. However, it is important to note that watershed approaches are prone to cause over-segmentation and often require extensive reliance on empirically tuned parameter settings.

Recent advancements in deep learning techniques provide promising possibilities for achieving more accurate cell segmentation performance. Many researchers have exploited representative networks including U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), and DeepLab (Chen et al., 2017) to segment and quantify cell morphological changes. Fabijańska (2018) trained the U-Net to differentiate pixels surrounding cell boundaries and skeletons, finally obtain the segmenation results via binarizing a boundary probability map. Vigueras-Guillén et al. (2019) introduced a local sliding window in UNet and generated probability labels to enhance the contrast between positive samples and background. Subsequently, they proposed a plug-and-play attention mechanism called feedback non-local attention to assist in inferring occluded cell regions (Vigueras-Guillén et al., 2022). Given the challenges of boundary discontinuity encountered when neural networks predict ambiguous cell boundaries, some studies considered combining the advantages of CNN and watershed. Lux and Matula (2020) integrated label-controlled watershed and convolutional networks to segment densely distributed cells, incorporating segmentation function criteria to describe object boundaries.

The CNN-based models are suitable for segmenting large cells, but for cells exhibiting artifacts within their bodies, complex post-processing algorithms are essential for separating cells that are in proximity, or for reconstructing fragmented cells to form a complete cellular structure. On the other hand, the segmentation performance of CNN decreases when facing cells of different sizes within the same field of view. With the popularity of Transformer (Vaswani et al., 2017), some studies have introduced Transformer with a global perspectives to support the segmentation process (Zhang et al., 2021; Zhu et al., 2022). Zhang et al. (2021) proposed a multi-branch hybrid transformer (MBT-Net) based on edge information, which utilized Transformer and residual connection to establish long-term dependencies between space and channels. Additionally, it also incorporated body edge branches to provide edge position. Zhu et al. (2022) designed a domain adaptive Transformer for atomy aware landmark detection for multi-domain learning. Oh and Jeong (2023) introduced a diffusion model-based data synthesis method aimed at mitigating variance among nuclear classes in tasks related to cell nucleus segmentation. To alleviate the learning bias caused by artificially designed disturbances in semi-supervised models, Zhou et al. (2023) proposed a consistency training method based on wavelet to address low-frequency and high-frequency information. Wang et al. (2023) introduced a two-stage knowledge distillation method designed to prevent the accumulation of errors resulting from noise artifacts.

Previous methods frequently employed Transformer to model dependency relationships among features within the layer of same size. Simultaneously, a feature within a specific layer only interacts directly with its adjacent feature layers, making it difficult to transmit hierarchical difference information of features. This poses a challenge to integrating multi-scale information from non adjacent layers at the macro scale. Our method leverages Transformer to model the hierarchical relationships among features across different layers, with the aim of reducing the deviation and loss of edge pixels caused by interpolation and sampling between features in different layers. We recursively convey context information across different feature layers within the structure of nnUNet. This approach allows for the acquisition of high-dimensional semantic relationships between pixel points and their neighbors from various perspective. Our contributions can be summarized as follows:

- We propose a Transformer-based network called U-NTCA to segment corneal stromal cell. It integrates with dual high-order channel attention and allows for the recursive transmission of crucial contextual features to better preserve detailed cell information.
- We introduce a high-order channel attention mechanism that extends the spatial interaction among pixels from second-order to higher-order. This procedure enables feature

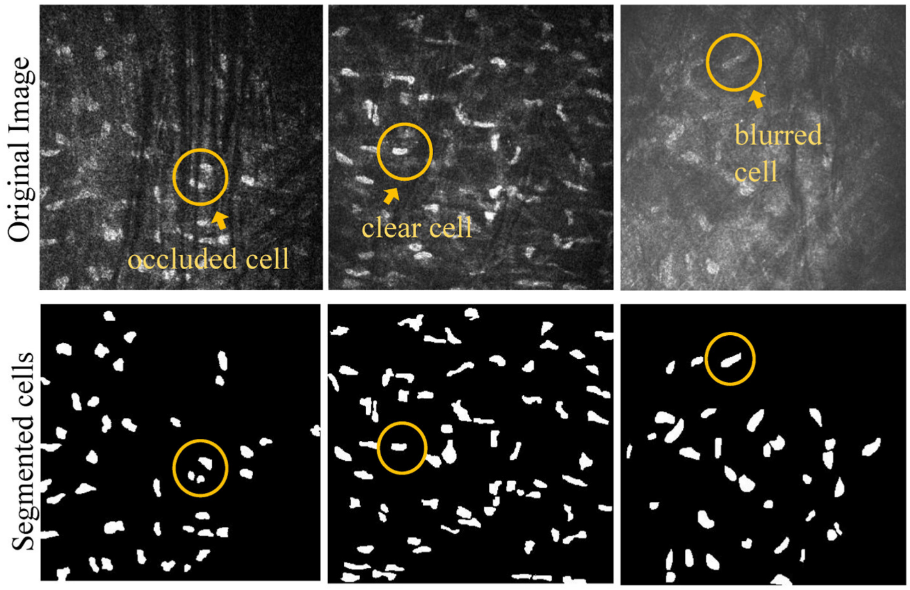

FIGURE 1
Example of different types of corneal stromal cells.

interaction within a low computational complexity by recursively increasing the channel width.
- We design a novel transformer-based method that combines a channel attention to generate multi-scale features. This facilitates direct feature transmission across non-adjacent layers in the network.

2 Dataset

All study subjects were scanned with a laser scanning corneal confocal microscopy HRTIII (Heidelberg Engineering, Heidelberg, Germany) at the affiliated Eye Hospital of Wenzhou Medical University. The study adhered to the tenets of the Declaration of Helsinki, and was approved by the Institutional Review Board of the Affiliated Eye Hospital of Wenzhou Medical University. All the participants provided a written informed consent after receiving an explanation of the risks/benefits of the study. The dataset utilized for this study on corneal stromal cells includes 136 images, each with a resolution of 384 × 384. The training dataset contains 96 images, while the testing dataset consists of 40 images. The segmentation labels of this dataset was manually annotated by one senior ophthalmologist using the ITK-SNAP software. During training, the data augmentation operations used in the training images include rotation, increasing contrast, adding noise, translation, and flipping. This dataset comprises corneal stromal cells source from three conditions: healthy corneas (named as "normal"), corneas with keratoconus (named as "cone"), and corneas that have been eroded by viruses (named as "HSK"). In general, these cells are presented in three different types. The first type exhibits a clear field of view and clear cell structure; The second type shows that the blood vessels in the background traverse the majority of visual field, causing partial occlusion of some corneal cells. The third type of image has severe blurriness, resulting in unclear cell edge morphology. Figure 1 shows typical examples of the corneal cell dataset. Table 1 provides a detailed description of the distribution of cells of different types in the test set.

TABLE 1 The distribution of different types of cells in the test set.

	Occlued cell	Blurred cell	Clear cell	All fields of view
Normal cell	4	4	16	24
Cone	2	2	6	10
HSK	4	0	2	6
All cell types	10	6	24	40

3 Methodology

3.1 nnUNet

In medical image segmentation, researchers often develop specific algorithms tailored to address distinct research tasks and solve targeted problems. This practice, however, can result in weak generalization and robustness for general models. nnUNet is proposed to specifically solve such issues of semantic segmentation tasks in medical imaging. It places a greater emphasis on aspects such as pre-processing, training, and post-processing procedures, with a primary focus on images. By systematically modeling various configuration strategies as a set of fixed parameters (learning rate and batch size), it proves adaptable to a range of medical image segmentation tasks.

The network architecture of nnUNet is the same as that of UNet, following the encoder-decoder paradigm, which comprises a series of dense convolutional blocks. Skip connections are employed between the encoder and decoder. By concatenating the generated features for use as complementary information, efficient feature mapping occurs between internal blocks, establishing convolutional and nonlinear connections. It is noteworthy that nnUNet, aiming to enhance stability and adaptability during training while avoiding limitations imposed by batch size, substitutes the original ReLU activation functions in UNet with leaky ReLUs. In addition, it replaces the more popular batch normalization with Instance normalization. This adaptation improves nnUNet with a stronger adaptive capability, effectively resolving training instability stemming from variations in imaging methods, sizes, and voxel spacing. This enables nnUNet to be employed across a variety of scenarios.

3.2 U-NTCA network

Considering nnUNet's outstanding data processing capability and parameter adaptive adjustment, we utilize it as a backbone network and enhance it to improve information interaction between pixels and the utilization of feature information. Figure 2 shows the overall structure of the proposed U-NTCA network. First, to highlight the relationship between neighboring pixels, our focus is on the three adjacent feature layers in the UNet. For the three feature layers with the same channel, we conduct feature dimension transformations on their heights and widths. The transformed outputs are used as inputs for the proposed g^nConv channel attention, facilitating higher-order operations. This process fosters efficient interaction between neighboring pixel regions. Subsequently, the enhanced features are integrated into the current aggregated features, which are then fed into the nested transformer to aid in generating full-resolution features. Additionally, the recursive transfer of underlying feature information mitigates ambiguity and reduces information loss resulting from the sampling process.

3.2.1 g^nConv high order attention mechanism

To enhance the interactive capabilities local context across varying resolutions, we introduce g^nConv module (Rao et al., 2022), which achieves explicit higher-order spatial interaction strategies within neighborhood. g^nConv is a module that implements channel attention through a combination of gated convolution and recursive strategy. It consists of three components: standard convolution, linear projections, and element-wise multiplications. It inherits the translation equivariant of standard convolution, thereby introducing inductive biases and avoiding the asymmetry arising from local attention.

Unlike the conventional approach of using g^nConv to directly interact with attention, we perform a morphological operation on feature $x_0 \in R^{H_0\times W_0\times C_0}$. This involves reshaping the dimensions of width and height $x \in R^{H\times W\times C}$, where $H = W = \sqrt{C_0}$ and $C = H_0 \times W_0$. This strategy aims to achieve high-order interaction between global pixels across diverse fields of view. It enables the network to learn the morphological characteristics and distribution patterns from varying perspectives and directions. For transformed feature x, we obtain mapping feature set $\phi_{in}(x)$ and feature auxiliary set $\left\{q_k\right\}_{k=0}^{n-1}$ with rich information embedding through the application of operation ϕ_{in}. The operation increases the feature dimension by two times, and then divides the expanded dimension according to rule C_k. It can be written as

$$[p_0^{HW\times C_0}, p_0^{HW\times C_0}, ..., q_{n-1}^{HW\times C_{n-1}}]$$

$$= \phi_{in}(x) \in R^{HW\times\left(C_0+\sum_{0\le k\le n-1} C_k\right)}$$

Subsequently, recursively execution of gated convolution is performed, introducing the interaction between adjacent features p_0 and q_0 through element-wise multiplications. This process achieves a spatial mixing input function with adaptive self-attention via

$$p_{k+1} = f_k\left(q_k\right) \odot g_k\left(p_k\right), \mathrm{k} = 0, 1,..., \mathrm{n} - 1$$

The channel dimension of each order can be written as

$$C_k = \frac{C}{2^{n-k-1}}, 0 \le \mathrm{k} \le \mathrm{n} - 1$$

Unlike the way that Transformer achieves spatial global interactions through mixing space tokens, g^nConv incrementally increases the channel width. It utilizes global computation of convolution and fully connected layers to expand the spatial interaction between pixels, progressing from second-order to higher-order interactions within less complexity.

3.2.2 Transformer nested with channel attention mechanism

In nnUNet, we transmit the features processed by g^nConv module as part of multi-scale features to Transformer. For downsampling image $x_d \in R^{H\times W\times d}$ and upsampling image $x_u \in R^{H\times W\times d}$, we flatten them to generate features $x_d \in R^{d\times HW}$ and $x_u \in R^{d\times HW}$. We utilize the g^nConv to encode x_u and generate $g^n(x_u)$ that interacts with neighboring pixels in a high-order space. Then, x_u, x_d, and $g^n(x_u)$ are sent to encoder to generate enhanced $\hat{x}_u$ through self-attention.

On one hand, the upsampling feature x_u is sent into the encoder, accompanied by its corresponding feature $g^n(x_u)$ that has undergone spatial point multiplication to facilitate higher-order interactions. This prompts the network to devote more attention to the decisive channels, implicitly reflecting the position of cells; On the other hand, $\hat{x}_u$ could bring more semantic information by fully interacting with the multi-scale features x_c transmitted from lower layers in decoder, guiding x_c to learn the constraint relationship between pixels and their neighbors from multi-scale perspectives. This aids in the inference of missing or incorrect cell regions caused by rough interpolation process. The specific formula for the attention mechanism is given as follows

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

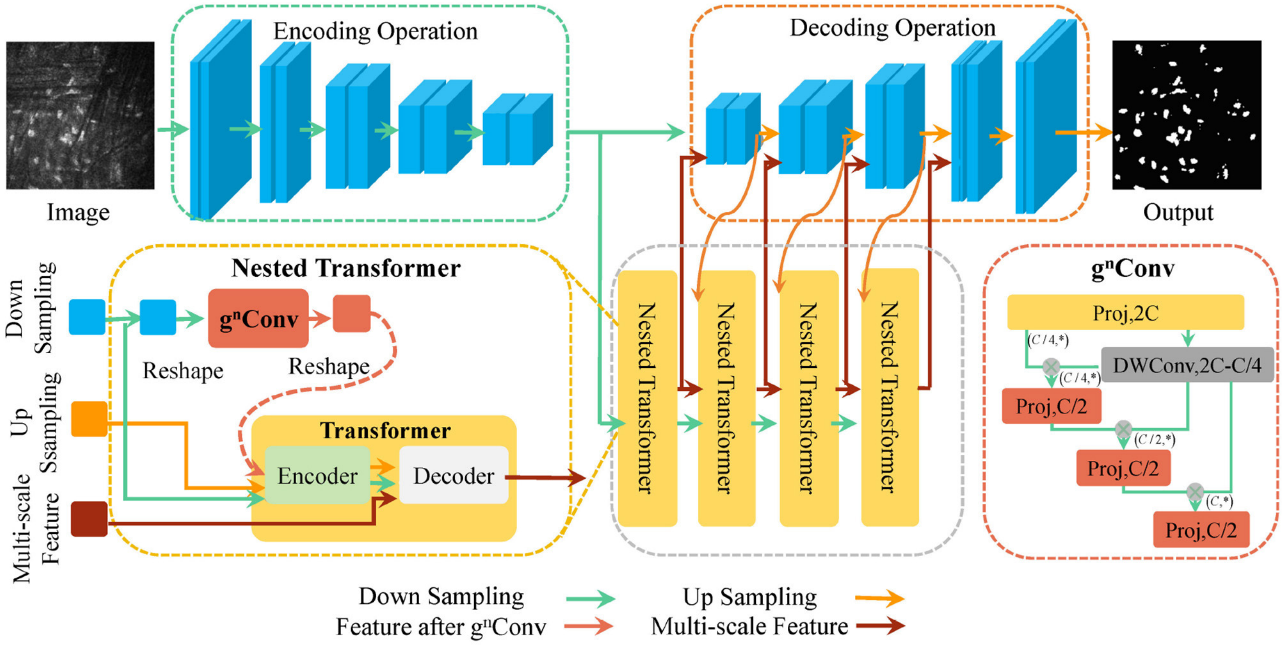

FIGURE 2
Schematic diagram of the proposed U-NTCA network.

The upsampling feature x_u, downsampling feature x_d, and enhanced feature $\hat{x}_u$ are encoded into $\hat{x}_u$. $\hat{x}_u$ contains information from higher-order pixel and their highly reliable distribution. At the same time, x_u also benefits from their attention interaction, creating conditions for comprehensive learning of the morphological structure and layout information of corneal cells in original image. The formula is written as follows

$$\hat{x}_u = x_u + Attention\left(x_u, x_d, g^n\left(x_u\right)\right)$$

Subsequently, the joint multi-scale feature x_c transmitted from lower layers is updated to $\hat{x}_c$ through the cross-attention mechanism. $\hat{x}_u$ and x_d collaboratively guide $\hat{x}_c$ in learning the potential mapping relationship between low-resolution targets and current targets of different scales. There is a size difference between the concatenated features transmitted from the bottom layer and the current layer features. We further feed the concatenated features transmitted from the bottom layer into the decoder to interact with the current layer features, exploring the implicit correspondence between downsampling and upsampling features between adjacent layers. We pass the concatenated multi-scale features as a medium for direct interaction among different layers. This approach facilitates the discrimination capability of ambiguous pixels. The formula is given as

$$\hat{x}_c = x_c + Attention\left(x_c, \hat{x}_u, x_d\right)$$

The $\hat{x}_c$ generated by x_c after cross attention is fed into the FFN (feedforward neural network) in residual form, which is a linear neural network with the following formula

$$FFN\left(\hat{x}_c\right) = \max\left(0, \hat{x}_c W_1 + b_1\right) W_2 + b_2$$

The process of generating $\tilde{x}_c$ through a FFN is written as

$$\tilde{x}_c = \hat{x}_c + FFN\left(\hat{x}_c\right)$$

We fuse the advanced multi-scale feature $\tilde{x}_c$ generated by decoder with upsampling feature of current layer in proportion to form $\tilde{x}_u$, providing more low-level local contextual information to the upsampling feature x_u that has information loss. $\tilde{x}_u$ is given by

$$\tilde{x}_u = \alpha x_u + (1 - \alpha)\,\tilde{x}_c$$

3.2.3 Recursive transmission of multi-scale features in U-shaped structures

We recursively implement the nested mechanism consisting of $g^n Conv$ and Transformer to deliver multi-scale features from different layers. Figure 3 displays the strategy of recursive transmission. In the process of generating upsampling features at full resolutions, we need to consider cascaded features transmitted from lower layers.

For the upsampling feature of the $i+1$ layer, its multi-scale feature x_u^{i+1} consists of the downsampling feature x_c^{i+1} of the i layer, the advanced encoding feature x_u^i, the decoding multi-scale feature x_d^i, $\tilde{x}_c^i$ and $g^n\left(x_u^i\right)$. Thus, $x_c^{i+1}(i > 1)$ is formulated as

$$x_c^{i+1} = H\left(x_d^i, \hat{x}_u^i, x_c^i, g^n\left(x_u^i\right)\right), \mathrm{i} = 2, 3, 4$$

For the lowest level features, the composition of its multi-scale features is illustrated in the following formula

$$x_c^{i+1} = H\left(x_d^i, \phi\left(x_u^i\right), \varphi(x_u^i), g^n\left(x_u^i\right)\right), \mathrm{i} = 1$$

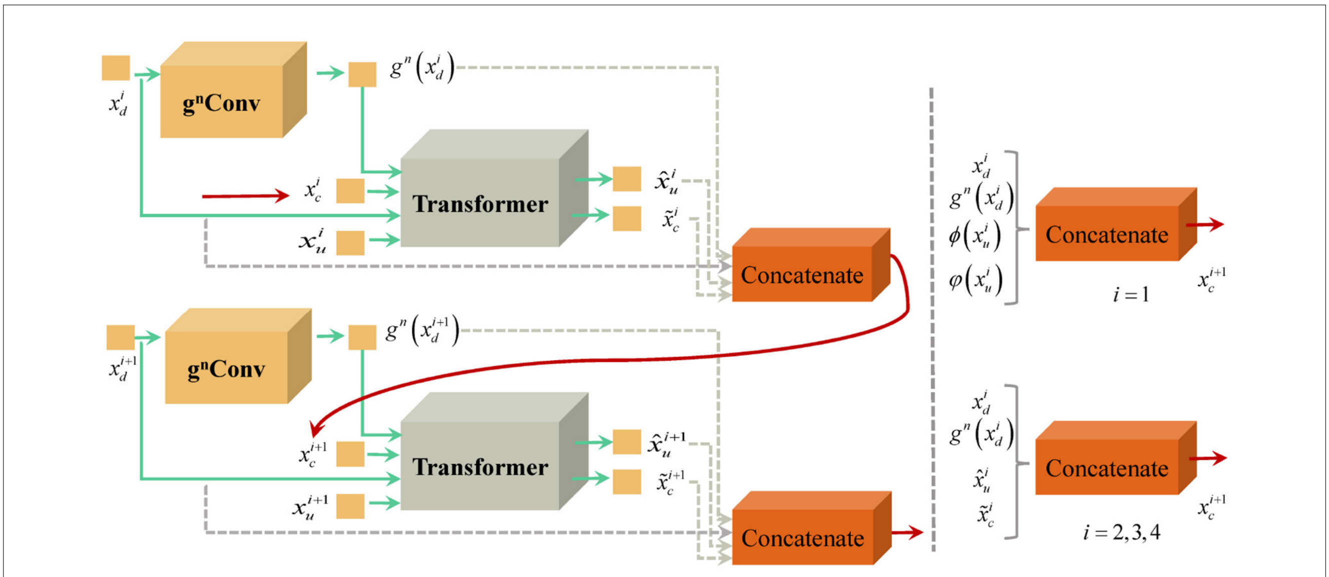

FIGURE 3
Schematic diagram of recursive transmission strategy.

with $\phi\,(x_u^i)$ and $\varphi\,(x_u^i)$ denote the intermediate steps in $g^n Conv$.

Although both $g^n Conv$ and nested Transformer leverage attention to improve cell pixel segmentation, they have inherent distinctions: (1) $g^n Conv$ attention operates at a high-dimensional level, facilitating information exchange among different channels. It processes a feature internally and allocates more attention to pivotal channels to obtain a optimal combination. This method enhances the ability to distinguish pixel positions and effectively filtering out background interference; (2) To enhance the network's ability to infer positional relationships among global features and similarities between features, the Nested Transformer are inserted at the bottleneck layer of the network and functions between different aggregated features. It is connected to the decoder during the upsampling phase, progressively propagating features from different scales. This results in obtaining distribution and layout constraints of corneal cells in a 2D plane, especially for challenging cells with weak luminance and blurred boundaries.

4 Experiments

4.1 Parameter settings

The experiments were conducted using PyTorch 1.7.1 on a GeForce RTX 3090 with 24GB of RAM. For the parameterization of $g^n Conv$, the number of iteration layers was set to $n = 3$, and the input features had a width (W) and height (H) of 22. The input feature channels followed the normal form rule 9×2^{2i} $(i = 1, 2, 3, 4)$. Regarding the converter network parameters, the overfitting value for the converter identification header was set to 0.1, and the forward feedback value was set to 2048. For the nested network features across different layers, the first three layers had 484 channels, and the fourth layer consisted of 256 channels. The training process employed a 5-fold cross validation method, further dividing the training and validation sets of the images in an 8:2 ratio. The fusion ratio of up-sampled features to corresponding multi-scale features was set to 3:7.

4.2 Evaluation metrics

In this experiment, we employ Dice, Acc, recall, pre (precision) and AUC as evaluation metrics to assess the segmentation performance. Dice quantifies the similarity between two samples, with values ranging from [0,1]. Pre (precision) denotes the proportion of correctly identified positive samples among all predicted positive samples, while recall represents the percentage of positive samples that were correctly predicted among all predicted samples. To clearly reflect the model's superior segmentation ability, Acc directly reflects the classification accuracy of the classifier. AUC quantifies the area under the ROC (Receiver Operating Characteristic) curve.

4.3 Comparative analysis

To verify the effectiveness of the proposed method, we compared the results of UNet++ (Zhou et al., 2018), Segformer (Xie et al., 2021), SwinUNet (Cao et al., 2022) and TransUNet (Chen et al., 2021) with the segmentation results of our method on the test set, as shown in Table 2. We can clearly observe that the proposed method outperforms other models in terms of all metrics. In Dice measure, the improved nnUNet reaches 82.71%, which was 23.35% higher than UNet++, 20.73% higher than Segformer, 10.85% higher than SwinUNet, 11.15% higher than TransUNet and 0.95% higher than nnUNet, respectively. Compared to nnUNet, the quantitative measurements of Dice, Acc, recall, pre, and AUC are improved by 0.08%, 0.62%, 0.55%, and 0.29%, respectively. It is demonstrated that our algorithm meets the requirement of accurate localization, thereby validating

TABLE 2 Comparison of experimental results of different encoding strategies for multi-scale features.

Method	Normal cell		HSK		Cone		All	
	Dice (%)	Acc (%)	Dice (%)	Acc (%)	Dice (%)	Acc (%)	Dice (%)	Acc (%)
UNet++	62.2	95.24	61.51	94.98	45.44	95.76	59.36	95.25
Segformer	65.02	95.49	60.34	94.97	52.55	95.97	61.98	95.43
SwinUNet	75.22	96.37	68.6	95.61	63.82	96.67	71.86	96.23
TransUNet	73.96	96.32	69.9	95.79	64.7	96.59	71.56	96.23
nnUNet	85.31	97.61	78.99	96.63	72.2	97.21	81.76	97.35
Our	**86.42**	**97.78**	**79.33**	**96.69**	**73.57**	**97.30**	**82.72**	**97.43**

Bold value indicates the best performance among all the methods in comparison.

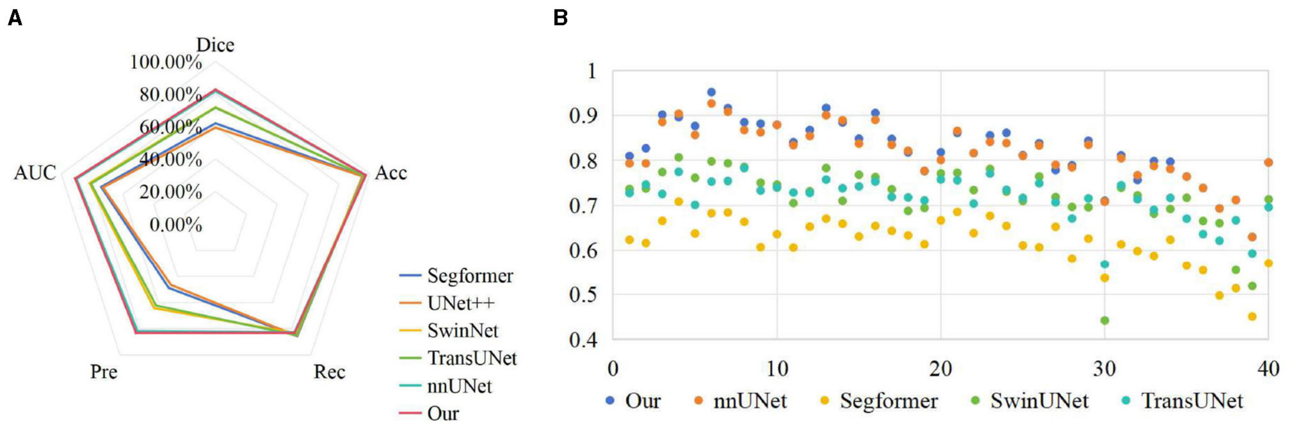

FIGURE 4
Comparison of visualization results of different methods on test set. **(A)** Comparison on various indicators; **(B)** Comparison on Dice index.

the effectiveness of the improved model. The results on the three classification datasets of Cell, HSK, and Cone intuitively show that our algorithm achieved the optimal performance on the Dice and Acc measures within these datasets. These results indicate that our method contributes comprehensively to the improvement of segmentation performance of nnUNet in multiple scenarios, rather than solving a single segmentation challenge alone. Figure 4A shows the comparison results of our method with other methods on different metrics, while Figure 4B shows the Dice values on the corneal test images of different methods. It can be intuitively seen that our method has achieved the best in all indicators, and at the same time, it outperforms other approaches in most of the test images.

4.4 Comparisons of different encoding strategies

As shown in Figure 5, we performed two comparative experiments to verify the influence of different encoding strategies of g^nConv and Transformer. To align with the dimension of high-level features, our method applied a concatenation on four smaller low-level features. In the comparative analysis in Table 3, we initially expanded the dimension of four low-level features via interpolation and then fused them with fixed proportional weights. The comparisons in Table 3 reveals that the concatenation strategy is superior to the interpolation strategy on most of the evaluation metrics. The multi-scale features based on concatenation achieve the performace of 82.72%, 97.43% and 83.06% respectively on Dice, Acc and recall. These values are respectively 0.29%, 0.11% and 2.51% higher than those achieved via interpolation. The above performance demonstrates the effectiveness of the concatenation strategy in conveying cell morphology and position distribution. This capability improves the localization of corneal cells with weaker contrast at upper layers, while the features generated through the interpolation strategy have certain information loss and ambiguous pixels, consequently diminishing the segmentation accuracy.

4.5 Ablation experiments

The ablation experiment in Table 4 verifies the impact of g^nConv and Transformer in the proposed framework. When leveraging only g^nConv information to enhance feature interactions, Dice and AUC were increased by 0.96% and 0.46%, respectively; Moreover, by incorporating a recursive Transformer into the U-shaped architecture of nnUNet, the improved model achieved Dice and AUC values of 82.72% and 90.93%, indicating further improvements accuracy. The experimental

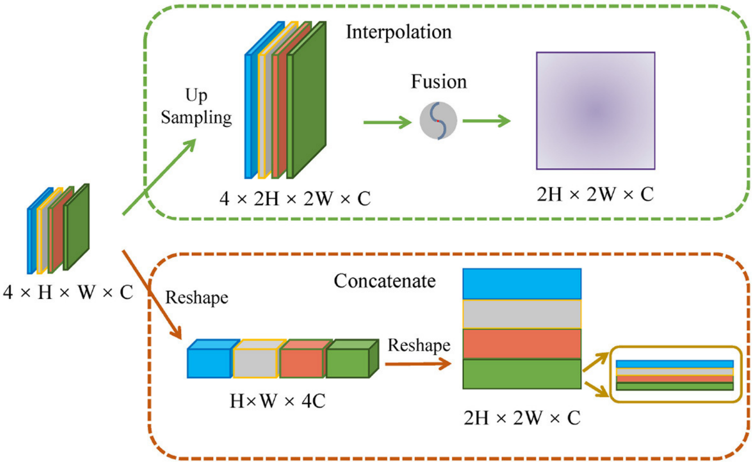

FIGURE 5
Schematic diagram of different encoding strategies.

TABLE 3 Comparison of experimental results of different encoding strategies for multi-scale features.

	Dice (%)	Acc (%)	recall (%)	Pre (%)
Interpolation	82.43	97.32	80.54	**85.20**
Concatenation	**82.72**	**97.43**	**83.06**	83.28

Bold value indicates the best performance among all the methods in comparison.

results demonstrate that the improved nnUNet model improves the results of Dice and AUC by 0.96% and 0.76% respectively, affirming the effectiveness of the proposed method.

4.6 Qualitative evaluation

As illustrated in Figure 6, a detailed visualization comparison is performed between nnUNet and our method on local image patches. In Patch 1 (a), nnUNet exhibits a larger area of false positives (magenta). In Patch 2 (a), nnUNet predicted more false positive cell parts compared to our method which has a more precise detection of cell boundary in patch 2 (b). The two cells in patch 3 belong to the challenge case of low visibility. Obviously, nnUNet missed one of the corneal stromal cells, while our method which is capable of detecting both of the cells. Figure 7 visualizes the heatmap of TransUNet, nnUNet, and our method. It can be intuitively seen that the TransUNet, which is designed based on Transformer, has less cells in warm colors (such as red and yellow) compared to the other two methods. However, it shows a significantly larger number of cells in cold colors (cyan and blue). In the heatmap of nnUNet, cells are predominantly warm-colored, with clear classification boundaries for positive and negative samples. The comparison between TransUNet and nnUNet highlights the distinction between CNN and Transformer. The latter focuses on the interaction between global context, and thus it performs better at identifying more cells (in cyan) that are difficult to recognize in a blurring condition. Our algorithm effectively combines the advantages of both approaches. As demonstrated in the two zoomed patches, our method not only has high predictive scores (with more red area) for the majority of cells in patch 1, but also successfully identifies a larger number of cells (in cyan) that were overlooked by the nnUNet in patch 2.

Figure 8 discusses the segmentation visualization results of different algorithms. In Image 1, the background vascular occlusion results in some intact cells being segmented into small fragments. nnUNet struggles to recognize some of the tiny cell fragments, whereas the proposed U-NTCA network successfully extracts the overall cell structures. Due to uneven illumination in Image 2, some cell edges are blurry with significant feature differences. This condition brings challenges for recognizing cells in dim illumination. Nevertheless, our method is able to detect more cells in low-contrast conditions. In Image 3, it can be observed that severe background interference obscures cell edges. Although the cells locate in areas with fair illumination, the accurate recognition of cell morphology and structure remains a challenging task. All the state-of-the-art approaches exhibit a notable disparity in achieving precise cell recognition, while nnUNet and our method outperforms the others in detecting a more complete cell contours.

5 Conclusion

The automatic and accurate segmentation of corneal stromal cells are essentially important to the rapid identification of abnormal lesions and timely prevention of the relevant diseases. To deal with the low segmentation accuracy of the existed methods under uneven illumination and occlusion, we designed a nested Transformer incorporated with nnUNet to model the implicit feature transmission across layers. The proposed model generates low-level positional and morphological features and are subsequently transmitted to upper layers to facilitate multi-scale feature fusion. In our future research, we intend to incorporate edge constraints to address challenges such as incorrectly connected cells or cells with broken edges. We will also further consider to establish a multi-task framework to achieve

TABLE 4 The ablation results of different modules.

	g^nConv	Transformer	Dice (%)	Acc (%)	Recall (%)	Pre (%)	AUC (%)
nnUNet	×	×	81.76	97.31	82.76	81.78	90.17
our	✓	×	82.16 (+0.40)	97.36 (+0.05)	82.44 (-0.32)	82.73 (+0.95)	90.63 (+0.46)
	✓	✓	**82.72 (+0.55)**	**97.43 (+0.07)**	**83.06 (+0.62)**	**83.28 (+0.55)**	**90.92 (+0.29)**

Bold value indicates the best performance among all the methods in comparison.

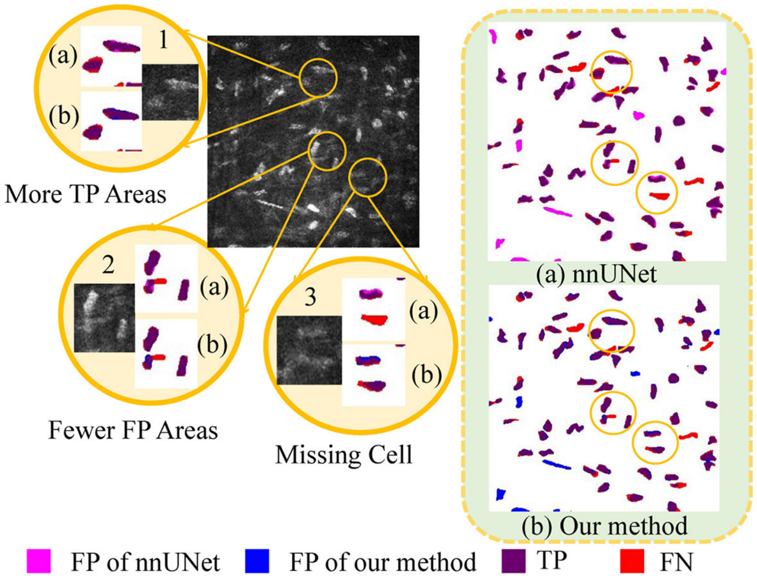

FIGURE 6
A detailed visualization comparison between nnUNet and our algorithm on local image patches.

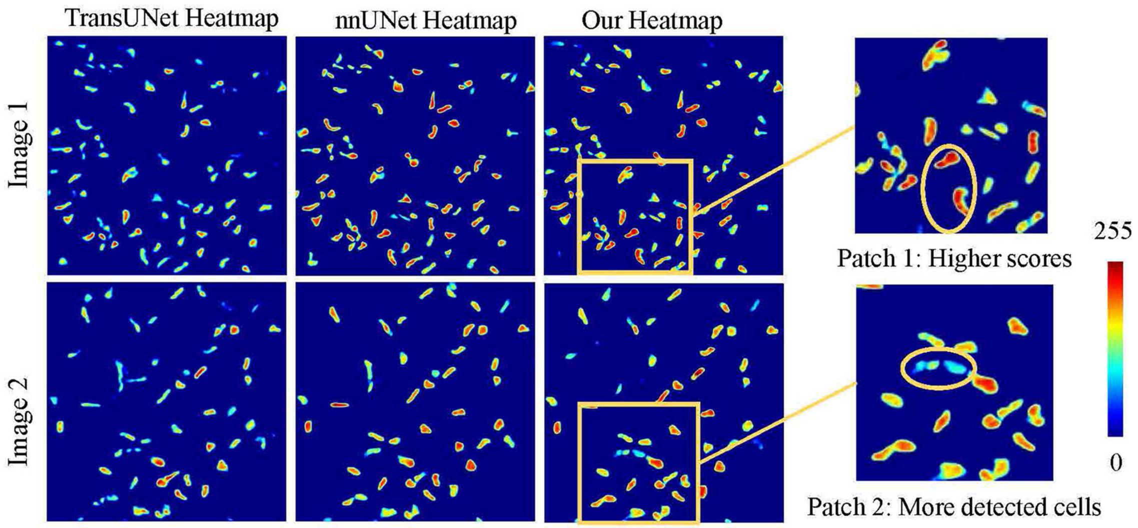

FIGURE 7
Heatmap visualization of different methods.

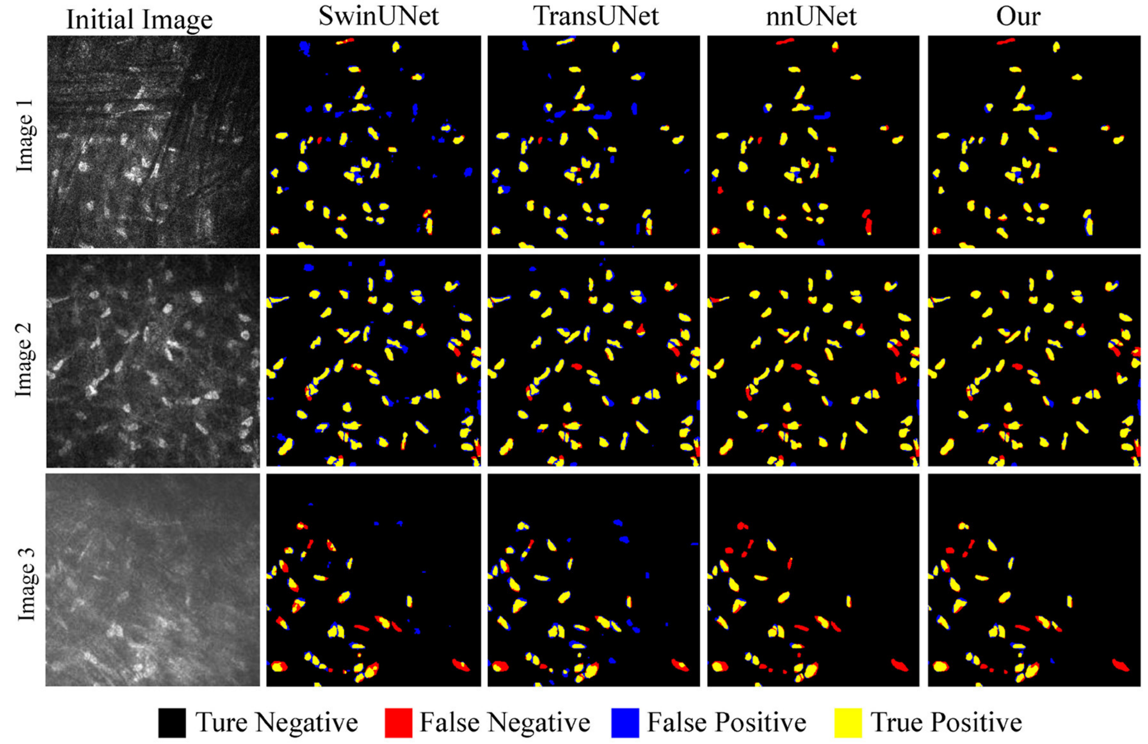

FIGURE 8
Example of overlapping results between different algorithms and real cell regions.

cell segmentation and diseases classification simultaneously, to promote computer-aided diagnosis.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Usage of the data should be under the permission of the corresponding authors. Requests to access these datasets should be directed to QZ, zhengqinxiang@aliyun.com.

Ethics statement

The studies involving humans were approved by the Affiliated Eye Hospital of Wenzhou Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

DZ: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. JinZ: Investigation, Validation, Writing – original draft. SL: Conceptualization, Data curation, Resources, Writing – review & editing. ZD: Data curation, Resources, Writing – review & editing. QZ: Data curation, Project administration, Supervision, Writing – review & editing. JioZ: Formal analysis, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the Zhejiang Provincial Natural Science Foundation (LQ23F010002 and LZ23F010002), in part by the Ningbo Natural Science Foundation (2022J143), and the Scientific Research Program of Zhejiang Provincial Department of Education (Y202250360), in part by the Research Startup Fund of Ningbo University of Technology (2022KQ29).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Fahdawi, S., Qahwaji, R., Al-Waisy, A. S., Ipson, S., Ferdousi, M., Malik, R. A., et al. (2018). A fully automated cell segmentation and morphometric parameter system for quantifying corneal endothelial cell morphology. *Comput. Methods Programs Biomed*. 160, 11–23. doi: 10.1016/j.cmpb.2018.03.015

Arıcı, C., Arslan, O. S., and Dikkaya, F. (2014). Corneal endothelial cell density and morphology in healthy turkish eyes. *J. Ophthalmol*. 2014, 852624. doi: 10.1155/2014/852624

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*. 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Barrientez, B., Nicholas, S. E., Whelchel, A., Sharif, R., Hjortdal, J., and Karamichos, D. (2019). Corneal injury: clinical and molecular aspects. *Exp. Eye Res*. 186, 107709. doi: 10.1016/j.exer.2019.107709

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2022). "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision* (Cham: Springer), 205–218.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). "Transunet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell*. 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Dagher, I., and El Tom, K. (2008). Waterballoons: a hybrid watershed balloon snake segmentation. *Image Vis. Comput*. 26, 905–912. doi: 10.1016/j.imavis.2007.10.010

Fabijańska, A. (2018). Segmentation of corneal endothelium images using a u-net-based convolutional neural network. *Artif. Intell. Med*. 88, 1–13. doi: 10.1016/j.artmed.2018.04.004

Lagali, N. (2020). Corneal stromal regeneration: current status and future therapeutic potential. *Curr. Eye Res*. 45, 278–290. doi: 10.1080/02713683.2019.1663874

Lux, F., and Matula, P. (2020). "Cell segmentation by combining marker-controlled watershed and deep learning," in *arXiv preprint arXiv:2004.01607*.

Oh, H.-J., and Jeong, W.-K. (2023). "Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets," in *arXiv preprint arXiv:2306.14132*.

Pan, Y., Zhou, T., and Xia, Y. (2015). "Bacterial foraging based edge detection for cell image segmentation," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE), 3873–3876.

Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S. N., and Lu, J. (2022). Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Adv. Neural Inf. Process. Syst*. 35, 10353–10366.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 234–241.

Shang, Z., Wang, X., Jiang, Y., Li, Z., and Ning, J. (2022). Identifying rumen protozoa in microscopic images of ruminant with improved yolact instance segmentation. *Biosyst. Eng*. 215, 156–169. doi: 10.1016/j.biosystemseng.2022.01.005

Sharif, J. M., Miswan, M., Ngadi, M., Salam, M. S. H., and bin Abdul Jamil, M. M. (2012). "Red blood cell segmentation using masking and watershed algorithm: a preliminary study," in *2012 international conference on biomedical engineering (ICoBE)* (Penang: IEEE), 258–262.

Subramaniam, M. D., Kumar, A., Chirayath, R. B., Nair, A. P., Iyer, M., and Vellingiri, B. (2021). Can deep learning revolutionize clinical understanding and diagnosis of optic neuropathy? *Artif. Intellig. Life Sci*. 1, 100018. doi: 10.1016/j.ailsci.2021.100018

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advance Neural Inf. Processing System*, 30.

Vigueras-Guillén, J. P., Sari, B., Goes, S. F., Lemij, H. G., van Rooij, J., Vermeer, K. A., et al. (2019). Fully convolutional architecture vs sliding-window cnn for corneal endothelium cell segmentation. *BMC Biomed. Engineer*. 1, 1–16. doi: 10.1186/s42490-019-0003-2

Vigueras-Guillén, J. P., van Rooij, J., van Dooren, B. T., Lemij, H. G., Islamaj, E., van Vliet, L. J., et al. (2022). Denseunets with feedback non-local attention for the segmentation of specular microscopy images of the corneal endothelium with guttae. *Sci. Rep*. 12, 14035. doi: 10.1038/s41598-022-18180-1

Wang, S., Zhang, D., Yan, Z., Shao, S., and Li, R. (2023). "Black-box source-free domain adaptation via two-stage knowledge distillation, in *arXiv preprint arXiv:2305.07881*.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst*. 34, 12077–12090.

Yan, M., Cai, J., Gao, J., and Luo, L. (2012). "K-means cluster algorithm based on color image enhancement for cell segmentation," in *2012 5th International Conference on BioMedical Engineering and Informatics* (Chongqing: IEEE),295–299.

Zhang, Y., Higashita, R., Fu, H., Xu, Y., Zhang, Y., Liu, H., et al. (2021). "A multi-branch hybrid transformer network for corneal endothelial cell segmentation," in *arXiv preprint arXiv:2106.07557*.

Zhou, Y., Huang, J., Wang, C., Song, L., and Yang, G. (2023). "Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images," in *Proceedings of the IEEE/CVF International Conference on Computer VisionI* (Montreal, BC: IEEE), 21085–21096.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018* (Granada: Springer), 3–11.

Zhu, H., Yao, Q., and Zhou, S. K. (2022). "Datr: Domain-adaptive transformer for multi-domain landmark detection," in *arXiv preprint arXiv:2203.06433*.