BMC Bioinformatics

RESEARCH

Open Access

DeShiftNet: a deformable-shifted cross-attention network for lightweight and robust organoid image segmentation

Le Tong[1], Tao Shu[1], Xinru Zhuang[1], Jingrui Bai[1], Lun Hu[2], Feng Tan[4], Yu-An Huang[5], Zhuhong You[5] and Pengwei Hu[2,3*]

*Correspondence:
Pengwei Hu
hpw@ms.xjb.ac.cn
[1]The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China
[2]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]AI and Quantum Lab, Merck KGaA, Darmstadt, Germany
[5]Northwestern Polytechnical University, Xi'an, China

Abstract

Background Organoid image segmentation is essential for quantitative analysis in disease modeling and drug screening, yet remains highly challenging due to substantial morphological variability and blurred boundaries in organoid images. Existing approaches often struggle to achieve a favorable balance between segmentation accuracy and computational efficiency.

Results In this paper, DeShiftNet, a lightweight segmentation framework, is proposed to extract discriminative features with high accuracy while maintaining low computational overhead. The model incorporates a deformable-shifted encoding strategy that adaptively samples local structures. It also includes a cross-attention-guided decoder for selective multi-scale feature alignment. Furthermore, a deformable multi-scale contextual refinement module enhances boundary coherence and contextual consistency. Extensive experiments on the multi-type OrganoID dataset show that DeShiftNet achieves competitive performance compared with recent segmentation models, while maintaining only 1.78M parameters and 2.65 GFLOPs. Notably, DeShiftNet achieves a Dice score of 0.961 on the Lung subset.

Conclusion These results indicate its potential practical value for efficient organoid segmentation in high-throughput experimental workflows.

Keywords Organoid image segmentation, Lightweight segmentation, Deformable-Shifted Encoding, Cross-Attention–Guided Decoder, Multi-scale contextual refinement

Introduction

Organoid models are three-dimensional in-vitro systems that recapitulate essential aspects of tissue architecture and function, and they have become indispensable for investigating organ development, tumor progression, and personalized therapeutic response [1–3]. Accurate segmentation of organoid microscopy images is fundamental for quantifying morphological characteristics, monitoring growth dynamics, and evaluating treatment effects [4, 5]. Nevertheless, pronounced morphological variability, optically induced boundary ambiguity, and heterogeneous texture patterns across specimens

render automated segmentation highly challenging [6]. These difficulties are further exacerbated in high-throughput imaging scenarios, where rapid and reliable inference must be achieved under limited computational resources.

Deep learning–based segmentation models have achieved remarkable progress in biomedical imaging, with encoder–decoder architectures such as U-Net and its variants [7–14] demonstrating strong capability in capturing multi-scale structures. However, convolution-based architectures rely on fixed receptive fields and local aggregation, which limits their ability to represent the irregular shapes and subtle boundary transitions commonly observed in organoid images. Transformer-based approaches [15–18] introduce global context modeling, yet their quadratic attention complexity leads to heavy computational costs and memory demands, making them less suitable for high-throughput or resource-constrained settings. Recent lightweight architectures, including token-shift MLP models such as UNeXt [19] and efficient decoders such as EMCAD [20], reduce parameters and FLOPs through compact feature mixing and attention mechanisms. Despite their effectiveness on general medical benchmarks, these models still struggle to fully adapt to heterogeneous organoid morphology and often fail to preserve fine structural details in blurred or low-contrast regions.

To address these limitations, we propose DeShiftNet, a lightweight segmentation framework specifically tailored for organoid imaging. The central idea is to enhance feature adaptivity and contextual reasoning while maintaining high computational efficiency. DeShiftNet employs a deformable-shifted encoding mechanism that adjusts spatial aggregation according to local structural cues, enabling more reliable representation of heterogeneous and curved organoid boundaries. A cross-attention–guided decoding strategy selectively aligns encoder and decoder features, facilitating the recovery of fine contours and interior textures that are often suppressed in lightweight networks. In addition, multi-scale contextual refinement strengthens information integration across resolutions, allowing the model to accommodate the wide range of organoid sizes and intensity variations. Through this design, DeShiftNet achieves accurate and stable organoid segmentation under realistic high-throughput constraints. The main contributions of this work are summarized as follows:

1. We propose a novel Deformable-Shifted Block (DSB) that introduces learnable spatial offsets into the token-shifting operation of UNeXt [19]. The DSB forms the core of the Deformable-Shifted Encoder (DSE), enabling content-aware deformation modeling and improving robustness to irregular organoid morphologies.
2. We design a Cross-Attention Gate (CAG) within the decoder to replace the grouped-convolution gate in EMCAD [20]. Unlike static convolutional gating, CAG computes channel-wise correlations between encoder and decoder features through lightweight convolutional projections, effectively enhancing multi-scale feature fusion and boundary recovery.
3. We incorporate a Deformable Multi-Scale Context Attention Module (D-MSCAM) into the decoder, which augments multi-branch convolutional modeling with a deformable convolutional sampling branch. This design allows adaptive aggregation of fine-grained spatial details and improves contextual consistency.
4. By combining deformable feature modeling (DSE and D-MSCAM) with cross-attention decoding (CAED), we construct DeShiftNet, an efficient end-to-end segmentation framework that achieves favorable accuracy and boundary delineation

performance on the OrganoID dataset [21] while maintaining extremely low computational complexity.

Related work

Medical image segmentation and organoid-specific challenges

Medical image segmentation serves as a cornerstone for quantitative biomedical analysis, and convolutional neural networks (CNNs) have long dominated this field. Classical architectures such as U-Net [7] and its variants—Unet++ [8], Atten-Unet [9], and R2U-Net [10]—combine multi-scale semantics with spatial detail through skip connections, achieving strong performance across diverse medical datasets [22]. However, their reliance on fixed receptive fields and static convolutional kernels limits their ability to capture long-range dependencies and adapt to complex spatial variations [23].

Organoid segmentation presents additional challenges beyond conventional organ-level imaging. As three-dimensional in-vitro tissue constructs, organoids exhibit pronounced intra-sample morphological heterogeneity, growth-driven structural changes, low-contrast or diffuse boundaries, and irregular texture patterns [24]. These characteristics often lead standard CNN architectures to produce unstable predictions, particularly when boundaries become blurred or when organoids deform or cluster densely. To alleviate these issues, several organoid-specific CNN models have been introduced. For example, OrganoID [21] builds up on U-Net with contour-aware post-processing to improve boundary preservation in bright-field organoid images, demonstrating improved robustness compared with classical CNN baselines. Nonetheless, CNN-based methods remain limited in capturing global context and adapting to curved or heterogeneous morphology.

To overcome the locality constraints of convolutions, Transformer-based architectures have been explored for biomedical segmentation. TransUNet first integrated Vision Transformers with CNN encoders to capture long-range dependencies [15], while Swin-Unet employed shifted-window attention to enhance hierarchical context modeling [18]. More recently, TransOrga [25] extended Transformer mechanisms to organoid imaging through multi-modal attention that merges structural priors across fluorescence and bright-field modalities. Despite their strong global reasoning capability, these models require substantial computational and memory resources due to self-attention's inherent quadratic complexity, making them less suitable for high-throughput organoid workflows. Furthermore, Transformer attention windows are typically fixed or window-local, limiting their sensitivity to fine-scale boundary cues that are critical for accurately delineating organoid contours.

Lightweight and adaptive architectures for efficient segmentation

Recent progress in lightweight segmentation architectures has focused on reducing computational cost while preserving accuracy, an essential requirement for high-resolution biomedical imaging. Models such as MobileNetV2-Unet and ESPNetv2 employ depthwise separable or factorized convolutions to reduce FLOPs and parameters, enabling fast inference in resource-constrained settings [26, 27]. However, these CNN-based designs often struggle to maintain fine structural fidelity when boundaries are weak or morphology varies markedly, as frequently observed in organoid microscopy.

Beyond convolutions, several efficient architectures incorporate token mixing or attention mechanisms to enhance global reasoning. UNeXt replaces convolutional bottlenecks with a Token-Shift MLP, achieving substantial model compression, but its fixed shift pattern limits adaptability to curved or heterogeneous organoid boundaries [19]. In the decoder domain, EMCAD introduces a multi-scale attention mechanism that improves feature fusion, yet its grouped large-kernel gating remains static and provides only coarse alignment across semantic levels, making it insufficient for restoring thin or low-contrast contours [20]. Adaptive modeling techniques—such as deformable convolutions [28], dynamic filters [29], and multi-branch context [30] modules—offer improved geometric flexibility, but they are typically applied only at isolated network stages or incur high computational overhead, constraining their practicality in lightweight medical segmentation.

Overall, existing efficient architectures lack a unified mechanism that simultaneously delivers content-adaptive spatial modeling, fine-grained encoder–decoder alignment, and multi-scale contextual reasoning—capabilities that are crucial for organoid segmentation characterized by irregular morphologies, blurry boundaries, and heterogeneous textures. These limitations motivate the design of DeShiftNet, which integrates deformable token interaction, cross-attention–guided fusion, and adaptive multi-scale refinement within a compact encoder–decoder framework tailored to organoid imaging.

Methodology

In this section, we introduce the proposed DeShiftNet segmentation framework and its core components. We first describe the overall network architecture and then detail each key module, including the Deformable-Shifted Encoder and Cross-Attention–Guided Decoder.

Overall network architecture

As illustrated in Fig. 1, DeShiftNet adopts a lightweight encoder–decoder architecture for efficient organoid image segmentation. The encoder starts with three U-Net–style convolutional stages, each consisting of two successive 3×3 convolutions with batch

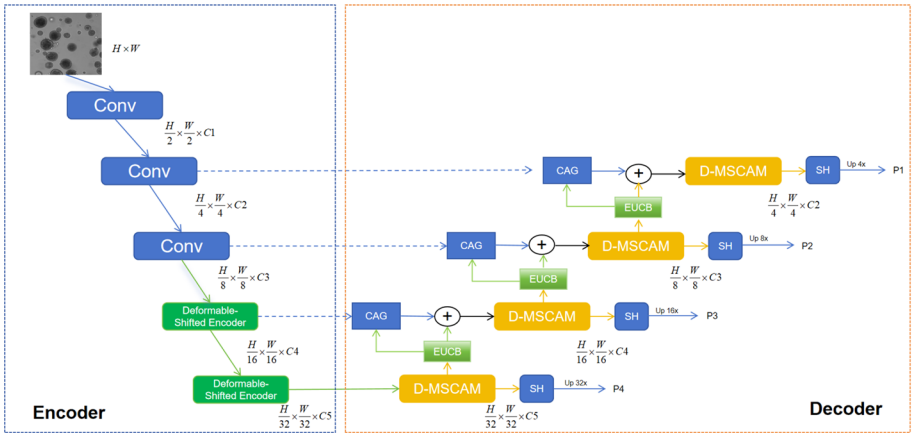

Fig. 1 Overall framework of DeShiftNet. The encoder consists of three convolutional stages followed by two Deformable-Shifted Encoder (DSE) stages, with channel dimensions C1 = 32, C2 = 64, C3 = 128, C4 = 160, and C5 = 256. The decoder upsamples features via Efficient Up-Convolution Blocks (EUCB), fuses encoder–decoder representations with Cross-Attention Gates (CAG), refines them using the Deformable Multi-Scale Context Attention Module (D-MSCAM), and generates multi-scale predictions P1–P4 through lightweight Segmentation Heads (SH)

normalization and GELU activation, followed by 2×2 max-pooling. These shallow layers provide stable texture and structure extraction while progressively reducing spatial resolution. In the deeper encoder, standard convolutions are replaced by the Deformable-Shifted Encoder (DSE), which introduces learnable spatial offsets into token mixing to enhance adaptability to heterogeneous organoid morphology. The decoder restores spatial resolution through hierarchical upsampling and feature fusion. The bottleneck feature is first refined by the Deformable Multi-Scale Context Attention Module (D-MSCAM), producing the coarsest auxiliary prediction P4. Features are then progressively upsampled by the Efficient Up-Convolution Block (EUCB); at each higher decoding stage, a Cross-Attention Gate (CAG) aligns encoder and decoder features before further D-MSCAM refinement, yielding intermediate predictions P3, P2, and P1. The final segmentation map is taken from P1 at one-quarter resolution, balancing spatial detail and computational cost.

Deformable-Shifted Encoder (DSE)

The Deformable-Shifted Encoder (DSE) constitutes the high-level feature extraction stage of DeShiftNet and is responsible for modeling complex organoid appearance variations in a lightweight yet adaptive manner. As illustrated in Figs. 1 and 2, the DSE receives the feature map from the third convolutional stage and first converts it into a token sequence through an Overlap Patch Embedding module, implemented as a 3×3 convolution with stride 2. The overlapping receptive fields moderately reduce spatial resolution while preserving local continuity, providing a structurally coherent token representation for subsequent deformable processing.

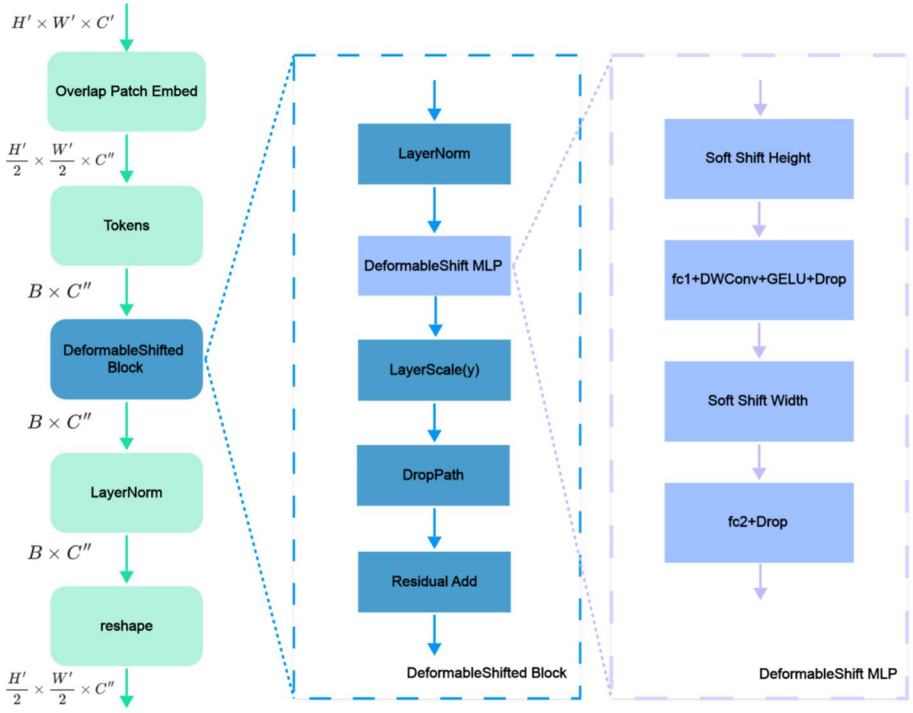

Fig. 2 Detailed architecture of the Deformable-Shifted Encoder in DeShiftNet, where the number of tokens B is $\frac{H'}{2} \times \frac{W'}{2}$

The transformed tokens are then fed into the Deformable-Shifted Block (DSB) module, whose internal structure is illustrated in the middle and right parts of Fig. 2. Inside the DSB, the input token sequence is first normalized by a LayerNorm layer. The normalized tokens are then passed to the DeformableShiftMLP, which serves as the core spatial-mixing unit. Unlike the fixed channel-shift pattern used in Shifted-MLP, DeformableShiftMLP performs a soft, content-adaptive shift along the spatial dimensions before lightweight mixing, enabling the effective receptive field to adapt to local structures.

For the height dimension, the shifted feature at position (i, j) is computed by aggregating cyclically shifted copies of the feature map with learned, spatially varying coefficients:

$$F_h(i,j) = \sum_{s=-p}^{p} \alpha_s(i,j)\mathrm{Roll}(F, s, \dim = 2), \tag{1}$$

where p denotes the maximum shift radius, $\alpha_s(i,j)$ are learned adaptive weights, and $\mathrm{Roll}(\cdot)$ performs cyclic shifting. An analogous operation is applied along the width dimension. This content-adaptive soft shift allows the effective receptive field to expand or contract according to local texture complexity, which is particularly beneficial for following curved boundaries, lumen-like structures, and gradual intensity transitions typical of organoid microscopy.

After the height-wise soft shift, the intermediate feature is processed by a lightweight mixing sequence composed of a linear projection, a depthwise convolution with GELU activation, and dropout. The resulting hidden representation is then reshaped back to a feature map and subjected to a second soft shift along the width dimension. It is finally projected by a second linear layer followed by dropout to produce the output of DeformableShiftMLP. To improve optimization stability and regularize training, we employ LayerScale together with DropPath, also known as Stochastic Depth [31], in the residual branch. Specifically, the token features are first normalized and processed by DeformableShiftMLP, then modulated by a learnable channel-wise scaling parameter, and finally added back to the shortcut branch through DropPath. With LayerScale and DropPath, the block output is formulated as:

$$T_{\text{out}} = T + \mathrm{DropPath}\big(\mathrm{LayerScale}\big(\mathrm{DeformableShiftMLP}(T)\big)\big), \tag{2}$$

where T is the input token sequence to the Deformable-Shifted Block and T_{out} denotes the residual-enhanced token output. Here, DropPath randomly drops the residual branch during training with a predefined probability, which acts as a regularization strategy to alleviate overfitting while preserving the full network depth at inference time. This design stabilizes deep representations and improves generalization without increasing inference cost. The output T_{out} is further normalized by LayerNorm and reshaped into a 2D feature map, which serves as the final output of the block.

Cross-attention–guided decoder

To ensure an efficient decoding pathway, we follow the lightweight multi-stage decoder of EMCAD proposed by Rahman et al. [20]. Specifically, each decoding stage retains the Efficient Up-Convolution Block (EUCB) as the upsampling backbone. EUCB restores spatial resolution via bilinear interpolation and then refines features with depthwise separable convolution and a 1×1 projection, which suppresses checkerboard artifacts

while keeping the computational overhead low. Built on this efficient upsampling skeleton, we introduce two targeted upgrades to overcome EMCAD's limitations on organoid imagery, namely the Cross-Attention Gate (CAG) and the Deformable Multi-Scale Context Attention Module (D-MSCAM). These components enhance encoder–decoder semantic alignment and boundary-aware contextual refinement, leading to more accurate multi-scale fusion and more reliable delineation of weak or low-contrast organoid boundaries.

Cross-Attention Gate (CAG)

EMCAD [20] introduced the Large-Kernel Grouped Attention Gate (LGAG) to fuse the decoder guidance g with the encoder detail feature x. However, because LGAG relies on a fixed large-kernel design, its receptive field cannot adapt to variations in local structures. This limitation makes it difficult to achieve accurate alignment of fine-scale features, which is especially unfavorable for organoid images where the boundaries are often weak, curved, and distributed across multiple spatial scales. To this end, we propose the CAG, which establishes finer channel-wise correlations through cross-attention. Given the decoder guidance feature g and the encoder detail feature x, CAG first performs lightweight 1×1 convolutional projections to a shared embedding space and then computes a fine-grained cross-correlation map via element-wise interaction. This correlation is further aggregated by a depthwise large-kernel convolution to encode local context before producing channel-aware gating weights $\boldsymbol{\psi}$ to modulate the encoder feature and obtain the fused output $\mathrm{x}_{\mathrm{out}}$. Formally, the gate is defined as:

$$\boldsymbol{\psi} = \sigma\left(W_{\boldsymbol{\psi}}\delta\left(\mathrm{DWConv}_{7\times7}\big(W_g(\mathrm{g}) \odot W_x(\mathrm{x})\big)\right)\right), \quad \mathrm{x}_{\mathrm{out}} = \mathrm{x} \odot \boldsymbol{\psi}, \tag{3}$$

where $W_g(\cdot)$, $W_x(\cdot)$, and $W_{\boldsymbol{\psi}}$ denote 1×1 convolutional projections, $\odot$ indicates element-wise multiplication, $\mathrm{DWConv}_{7\times7}$ is a depthwise 7×7 convolution for context aggregation, $\delta(\cdot)$ is a lightweight nonlinearity with normalization, and $\sigma(\cdot)$ is the sigmoid activation. Compared with LGAG, the multiplicative cross-correlation explicitly measures feature consistency between encoder and decoder at each spatial location and channel, enabling more precise multi-scale alignment and better preservation of thin and low-contrast organoid boundaries, while keeping the gate computationally compact.

Deformable Multi-Scale Context Attention Module (D-MSCAM)

The Deformable Multi-Scale Context Attention Module (D-MSCAM) is developed by extending the MSCAM decoder in EMCAD with deformable, content-adaptive context modeling, aiming to better handle the irregular morphology and weak boundary contrast typical of organoid microscopy. As shown in Fig. 3, the input decoder feature is first reweighted by a Channel Attention Block (CAB) and a Spatial Attention Block (SAB), so that informative channels and salient spatial regions are emphasized before multi-scale context aggregation [20]. The CAB applies adaptive average pooling and adaptive max pooling over the spatial dimensions to aggregate global channel cues, followed by lightweight channel reduction and restoration, and then sigmoid-based gating to recalibrate channel responses. The SAB computes spatial attention by performing channel-wise average pooling and max pooling, followed by a large-kernel convolution and sigmoid activation to generate spatial attention weights.

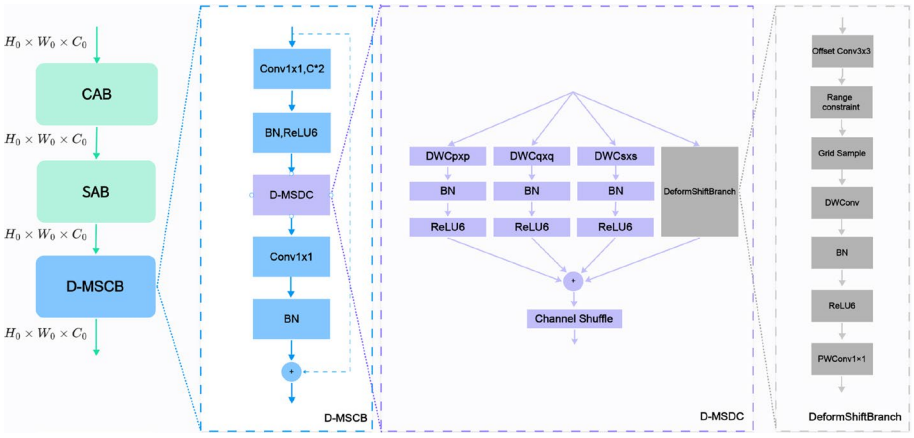

Fig. 3 Detailed architecture of the Deformable Multi-Scale Context Attention Module (D-MSCAM) in DeShiftNet

The reweighted feature is then processed by the Deformable Multi-Scale Convolution Block (D-MSCB), whose core operator is the Deformable Multi-Scale Depthwise Convolution (D-MSDC). D-MSDC preserves the efficient parallel depthwise branches with fixed receptive fields to capture scale specific patterns, and further introduces a deformable shifted token branch to enable dynamic sampling along organoid boundaries. Specifically, the deformable branch predicts spatial offsets from the input feature and performs content adaptive warping as:

$$\boldsymbol{x}^{\mathrm{def}} = \mathrm{GridSample}\big(\boldsymbol{x}, \boldsymbol{g} + \tanh\big(\mathrm{Conv}_{3\times3}(\boldsymbol{x})\big) \cdot s\big) \tag{4}$$

where $\boldsymbol{x}$ denotes the input feature to the deformable branch, $\boldsymbol{g}$ is the regular sampling grid, $\mathrm{Conv}_{3\times3}(\cdot)$ is a 3×3 convolution used to regress offsets, $\tanh(\cdot)$ constrains each offset element to the range $[-1, 1]$, s is a scalar controlling the maximum displacement magnitude, and $\mathrm{GridSample}(\cdot)$ represents bilinear grid sampling.

This content adaptive warping modulates the effective receptive field according to local appearance, allowing the module to follow curved organoid contours and alleviate boundary deviations caused by local blur or shape variability. The warped feature is then refined by a lightweight depthwise convolution followed by normalization, GELU activation, and pointwise projection to form the deformable branch response. In parallel, depthwise branches with kernels 1×1, 3×3, and 5×5 capture complementary context at different spatial extents. Their outputs are fused with the deformable response and mixed via channel shuffle, and the aggregated representation is finally added back through a residual connection, preserving the original decoder signal while injecting multi-scale, content adaptive context that benefits organoid images with heterogeneous morphology and weak, ambiguous boundaries.

Segmentation heads with multi-scale deep supervision and hybrid loss

To produce pixel-wise predictions at different decoding stages, DeShiftNet attaches an independent Segmentation Head (SH) to each decoder output, as illustrated in Fig. 1. Each SH is intentionally lightweight, implemented as a 1×1 convolution followed by a Sigmoid activation, thereby providing a direct projection from feature space to probability space with negligible computational overhead. Let P_i denote the probability map generated at the i-th decoding stage and Y the corresponding ground truth mask. The

per-stage supervision adopts a hybrid objective combining Binary Cross-Entropy (BCE) and Dice losses [32]:

$$\mathcal{L}_{\mathrm{BCE}}(P_i, Y) = -\frac{1}{N}\sum_{j=1}^{N}\left[y_j \log p_{i,j} + (1-y_j)\log(1-p_{i,j})\right], \tag{5}$$

$$\mathcal{L}_{\mathrm{Dice}}(P_i, Y) = 1 - \frac{2\sum_{j=1}^{N} y_j p_{i,j} + \epsilon}{\sum_{j=1}^{N} y_j + \sum_{j=1}^{N} p_{i,j} + \epsilon}, \tag{6}$$

$$\mathcal{L}(P_i, Y) = \alpha\mathcal{L}_{\mathrm{BCE}}(P_i, Y) + (1-\alpha)\mathcal{L}_{\mathrm{Dice}}(P_i, Y), \tag{7}$$

where N is the number of pixels, $p_{i,j} \in [0,1]$ is the predicted probability at pixel j for stage i, $y_j \in \{0,1\}$ is the corresponding binary label, ϵ is a small constant for numerical stability, and $\alpha \in [0,1]$ balances region fidelity and overlap consistency (set to α=0.5 in our experiments).

During training, multi-scale deep supervision is enforced on the stage outputs $\{P1, P2, P3, P4\}$. All decoding stages contribute equally to the objective, and an auxiliary loss is further applied to the fused prediction obtained by summing all stage outputs. The overall training loss is therefore defined as:

$$\mathcal{L}_{\mathrm{total}} = \sum_{i=1}^{4}\lambda_i\mathcal{L}(P_i, Y) + \lambda_0\mathcal{L}\left(\sum_{i=1}^{4} P_i, Y\right), \tag{8}$$

where λ_i=1 (i=0, ..., 4). This deeply supervised hybrid objective stabilizes gradient propagation across scales and encourages consistent boundary localization. At inference time, only the highest-resolution prediction P_1 is retained as the final segmentation output.

Experiments

Datasets

All experiments were conducted on the publicly released OrganoID dataset [21], a benchmark for organoid segmentation that covers diverse tissue origins and morphological characteristics. The dataset contains brightfield and phase-contrast microscopy images of organoids, with the training images derived from PDAC organoids from two different patients cultured either on a standard tissue culture plate or on a microfluidic organoid platform. The benchmark includes 66 manually annotated PDAC organoid images, among which 52 images were used for training-set construction and the remaining 14 images were reserved for validation. The publicly released training set used in this study consists of 2000 augmented images generated from the 52 original training images using the Augmentor library, with augmentation operations including random rotation (up to $\pm 20°$), horizontal and vertical flipping, random zooming (percentage area = 0.7), shearing (up to $\pm 20°$), random distortion (grid width/height = 5, magnitude = 3), skewing (magnitude = 0.3), and resizing to 512×512 pixels. This augmented training set increases sample diversity while aiming to preserve the overall morphology and structural characteristics of organoids.

To further assess the model performance under diverse organoid morphologies and imaging characteristics, an additional test set of 28 images was used, including PDAC (10

images), adenoid cystic carcinoma (ACC, 6 images), colon epithelium (Colon, 6 images), and distal lung epithelia (Lung, 6 images). These test images exhibit substantial variations in organoid density, size distribution, and background complexity, thereby posing considerable challenges for segmentation. Representative examples of the released augmented training images and the test images are provided in Fig. 4, allowing readers to better assess the differences between the training and test samples.

For network input standardization, all images and their corresponding masks were resized to a uniform spatial resolution of 512×512. The input images were then normalized before being fed into the network, while the ground-truth masks were encoded in binary format.

Implementation details

The experiments were implemented in PyTorch on an NVIDIA RTX A16 GPU with 15GB memory. The AdamW optimizer was employed with an initial learning rate of 1×10^{-4}, decayed to 1×10^{-6} via cosine annealing scheduling. The batch size was set to 8, and the model was trained for 500 epochs. Model performance is evaluated using standard medical image segmentation metrics, including Precision, Recall, mean Intersection-over-Union (mIoU), and Dice coefficient—the latter being widely adopted due to its sensitivity to foreground overlap. Furthermore, to comprehensively assess practical deployability, we report the number of trainable parameters (Params) and floating-point operations (FLOPs), which reflect model complexity and computational cost, respectively. In addition, inference speed in frames per second (FPS) and GPU memory consumption are also reported to provide a more practical evaluation of computational efficiency under the same hardware setting. This lightweight architecture significantly reduces resource requirements, making it suitable for high-throughput biomedical analysis scenarios.

Quantitative and qualitative comparison

To ensure a comprehensive and fair evaluation of DeShiftNet on organoid segmentation, we compare it with a diverse set of state-of-the-art segmentation models representing different architectural paradigms in medical imaging. Atten-Unet [9] introduces attention gates to suppress irrelevant responses in traditional U-Net. BCS-Net [33] leverages boundary, context, and semantic reconstruction to enhance segmentation of complex medical structures. GFNet [34] employs dynamic global filtering to improve the representation of small-scale and fine-grained details. ACC-Unet [35] adopts a fully convolutional U-Net design tailored for organoid-like structures with irregular shapes.

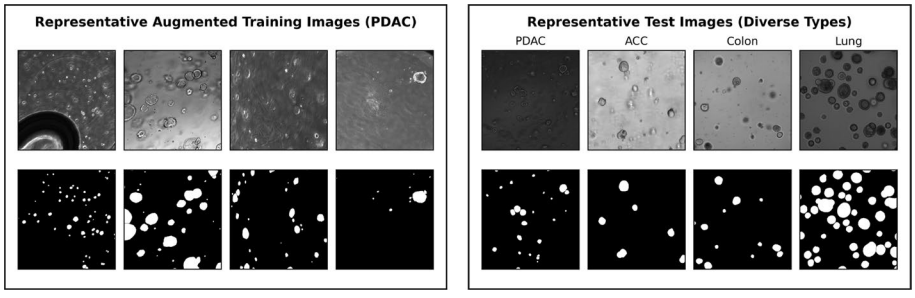

Fig. 4 Representative examples of the augmented training set and the diverse test set for organoid segmentation

Swin-Unet [18] and TransUNet [15] incorporate Transformer-based global modeling, enabling long-range context aggregation. Cellpose [36] is a widely used generalist segmentation framework for biological images and serves as an important practical baseline in biomedical image analysis. MultiFusionNet [37] utilizes multi-scale hierarchical feature fusion to strengthen robustness in challenging conditions. UNeXt [19] represents an efficient MLP-based design that achieves competitive performance with very low parameter cost. OrganoID [21] is a lightweight organoid-specific model widely used as a baseline in organoid segmentation tasks. PVT-EMCAD-B0 combines Pyramid Vision Transformer encoding with the EMCAD decoder for efficient multi-scale feature decoding [20]. Together, these baselines cover CNN-based, attention-enhanced, Transformer-driven, MLP-based, generalist biological image segmentation, and organoid-specific models, providing a broad and rigorous benchmark to validate the effectiveness of our proposed DeShiftNet in terms of both segmentation accuracy and computational efficiency.

In Tables 1, 2, 3, values in bold indicate the optimal results in each comparison. As shown in Table 1, DeShiftNet achieves the best performance among all compared methods in terms of Dice, mIoU, Precision, and Recall on the PDAC organoid dataset, outperforming both large models such as TransUNet [15] and lightweight baselines including UNeXt [19] and OrganoID [21]. This performance advantage is closely related to several architectural components that match the characteristics of organoid images. The Deformable-Shifted Encoder improves feature extraction in regions with weak boundaries and irregular shapes, the Cross-Attention Gate enhances feature alignment during decoding, and the D-MSCAM module strengthens contextual modeling under complex scale and intensity variations. With the help of multi-scale deep supervision, these components enable DeShiftNet to better handle the blurred textures, heterogeneous morphologies, and fine structural details commonly observed in organoid imagery, leading to more accurate segmentation under a lightweight design. Although DeShiftNet does not attain the minimum parameter count or the highest inference speed among all compared methods, it maintains a favorable efficiency profile, with 1.78M parameters, 2.65G FLOPs, 52.8 FPS, and 204 MB GPU memory consumption, indicating a balanced trade-off between segmentation accuracy and computational cost.

To further visualize the trade-off between accuracy and efficiency, Fig. 5 presents a three-dimensional plot of Params, FLOPs, and Dice on the PDAC dataset. The x-axis

Table 1 Quantitative comparison of segmentation performance on the PDAC organoid dataset

PDAC								
Model	**#Params**	**#FLOPs**	**FPS**	**GPU Mem**	**Dice**	**mIoU**	**Precision**	**Recall**
Atten-Unet	34.88M	266.56G	12.81	1456 MB	0.789	0.691	0.749	0.782
BCS-Net	44.82M	59.36G	7.52	2677 MB	0.782	0.658	0.734	0.805
Swin-Unet	27.17M	32.39G	14.81	942 MB	0.783	0.635	0.738	0.773
TransUNet	105.32M	154.08G	3.68	4580 MB	0.847	0.695	0.812	0.852
Cellpose	6.62M	20.43G	16.34	458 MB	0.865	0.768	0.846	0.889
GFNet	18.10M	107.54G	20.74	1228 MB	0.805	0.668	0.787	0.796
ACC-Unet	16.80M	198.51G	12.85	5738 MB	0.788	0.662	0.784	0.823
MultiFusionNet	7.30M	5.75G	25.7	896 MB	0.801	0.675	0.773	0.847
UNeXt	1.47M	**2.29G**	**64.9**	**134** MB	0.851	0.740	0.840	0.802
OrganoID	**0.49M**	6.93G	5.46	242 MB	0.859	0.752	0.702	0.836
PVT-EMCAD-B0	3.92M	3.36G	40.5	312 MB	0.838	0.726	0.781	0.920
Ours	1.78M	2.65G	52.8	204 MB	**0.887**	**0.798**	**0.876**	**0.922**

Table 2 Quantitative segmentation results on the ACC, Colon and Lung organoid datasets

Model	#Params	#FLOPs	FPS	GPU Mem	Dice	mIoU	Precision	Recall
ACC								
Atten-Unet	34.88M	266.56G	12.81	1456 MB	0.782	0.662	0.687	0.875
BCS-Net	44.82M	59.36G	7.52	2677 MB	0.815	0.672	0.706	0.881
Swin-Unet	27.17M	32.39G	14.81	942 MB	0.808	0.682	0.682	0.876
TransUNet	105.32M	154.08G	3.68	4580 MB	0.834	0.721	0.755	0.879
Cellpose	6.62M	20.43G	16.34	458 MB	0.851	0.763	0.865	0.884
GFNet	18.10M	107.54G	20.74	1228 MB	0.826	0.703	0.731	0.885
ACC-Unet	16.80M	198.51G	12.85	5738 MB	0.819	0.697	0.728	0.867
MultiFusionNet	7.30M	5.75G	25.7	896 MB	0.828	0.714	0.781	0.892
UNeXt	1.47M	**2.29G**	**64.9**	**134** MB	0.802	0.670	0.880	0.844
OrganoID	**0.49M**	6.93G	5.46	242 MB	0.848	0.736	0.622	0.866
PVT-EMCAD-B0	3.92M	3.36G	40.5	312 MB	0.839	0.727	0.737	0.892
Ours	1.78M	2.65G	52.8	204 MB	**0.902**	**0.823**	**0.899**	**0.908**
Colon								
Atten-Unet	34.88M	266.56G	12.81	1456 MB	0.830	0.743	0.718	0.755
BCS-Net	44.82M	59.36G	7.52	2677 MB	0.817	0.721	0.719	0.745
Swin-Unet	27.17M	32.39G	14.81	942 MB	0.833	0.750	0.709	0.738
TransUNet	105.32M	154.08G	3.68	4580 MB	0.852	0.775	0.744	0.766
Cellpose	6.62M	20.43G	16.34	458 MB	0.878	0.789	0.866	0.887
GFNet	18.10M	107.54G	20.74	1228 MB	0.845	0.751	0.736	0.751
ACC-Unet	16.80M	198.51G	12.85	5738 MB	0.836	0.744	0.722	0.750
MultiFusionNet	7.30M	5.75G	25.7	896 MB	0.873	0.776	0.899	0.851
UNeXt	1.47M	**2.29G**	**64.9**	**134** MB	0.868	0.767	0.801	0.890
OrganoID	**0.49M**	6.93G	5.46	242 MB	0.867	0.766	0.674	0.850
PVT-EMCAD-B0	3.92M	3.36G	40.5	312 MB	0.823	0.704	0.721	0.892
Ours	1.78M	2.65G	52.8	204 MB	**0.904**	**0.826**	**0.896**	**0.913**
Lung								
Atten-Unet	34.88M	266.56G	12.81	1456 MB	0.902	0.868	0.855	0.929
BCS-Net	44.82M	59.36G	7.52	2677 MB	0.912	0.862	0.878	0.923
Swin-Unet	27.17M	32.39G	14.81	942 MB	0.911	0.851	0.872	0.929
TransUNet	105.32M	154.08G	3.68	4580 MB	0.928	0.879	0.902	0.941
Cellpose	6.62M	20.43G	16.34	458 MB	0.926	0.874	0.915	0.948
GFNet	18.10M	107.54G	20.74	1228 MB	0.910	0.848	0.897	0.932
ACC-Unet	16.80M	198.51G	12.85	5738 MB	0.922	0.870	0.883	0.927
MultiFusionNet	7.30M	5.75G	25.7	896 MB	0.935	0.877	0.939	0.931
UNeXt	1.47M	**2.29G**	**64.9**	**134** MB	0.937	0.882	0.907	0.968
OrganoID	**0.49M**	6.93G	5.46	242 MB	0.911	0.836	0.794	0.938
PVT-EMCAD-B0	3.92M	3.36G	40.5	312 MB	0.934	0.876	0.882	0.956
Ours	1.78M	2.65G	52.8	204 MB	**0.961**	**0.924**	**0.955**	**0.979**

Table 3 Ablation results on the OrganoID dataset

Model	Dice	mIoU	Precision	Recall
Ours (Baseline + DSE + CAG + D-MSCAM)	**0.909**	**0.836**	**0.902**	**0.969**
Ours w/ Standard Shift	0.877	0.798	0.875	0.924
Ours w/o DSE	0.862	0.762	0.848	0.913
Ours w/o CAG	0.873	0.781	0.866	0.902
Ours w/o D-MSCAM	0.880	0.792	0.871	0.893
Ours w/o CAG and D-MSCAM	0.830	0.738	0.784	0.926
Baseline	0.779	0.681	0.716	0.899

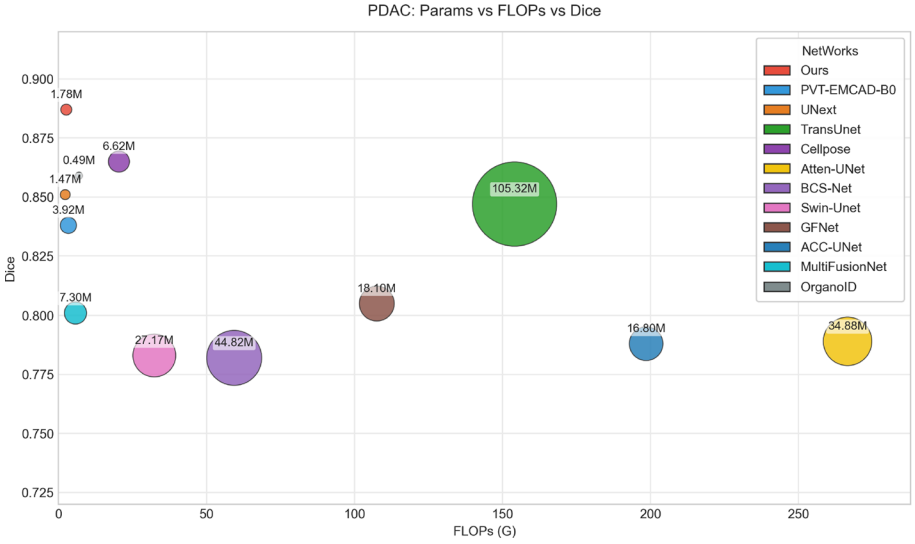

Fig. 5 Comparison of model performance on the PDAC dataset in terms of Dice score versus computational cost. The horizontal axis represents FLOPs (in G), and the vertical axis denotes Dice coefficient, while bubble size indicates the number of parameters

represents FLOPs (in G), the y-axis shows the Dice, and bubble size indicates Params. As observed, DeShiftNet (red dot) lies in a favorable region of the accuracy-efficiency trade-off space, achieving the highest Dice while maintaining relatively low model complexity and computational cost.

To validate generalization capability, we directly apply DeShiftNet—trained solely on PDAC data—to three unseen organoid types (ACC, Colon, Lung) without any fine-tuning. Results in Table 2 show that DeShiftNet achieves a Dice of 0.902 on ACC, outperforming OrganoID (0.848) and Cellpose (0.851); 0.904 on Colon, significantly exceeding PVT-EMCAD-B0 (0.823) and MultiFusionNet (0.873); and an impressive 0.961 on Lung, surpassing all baselines including TransUNet (0.928) and GFNet (0.910). This indicates that the dynamic shift-based modeling in DeShiftNet effectively captures shared morphological priors across organoid types, suggesting promising transferability across the evaluated organoid subsets.

Building upon the quantitative assessment, we now turn to qualitative insights provided by Fig. 6, which presents representative segmentation results of DeShiftNet and the compared methods across four organoid types: ACC, Lung, Colon, and PDAC. These samples exhibit substantial variation in morphology and imaging characteristics, including small and sparsely distributed targets, densely packed structures, ambiguous boundaries, and fine-scale details. Despite these challenges, DeShiftNet generally produces predictions that are more consistent with the ground truth than those of the compared methods.

Nevertheless, Fig. 7 presents a representative failure case from the ACC test set. In this example, low contrast, blurred background interference, and closely adjacent organoids with ambiguous boundaries make accurate delineation particularly difficult. Although DeShiftNet still captures the major foreground regions, some small organoids are missed, several clustered structures are merged into larger connected predictions, and slight boundary deviations remain for irregularly shaped targets. These errors reflect not only the intrinsic difficulty of the image, but also the limitations of the current

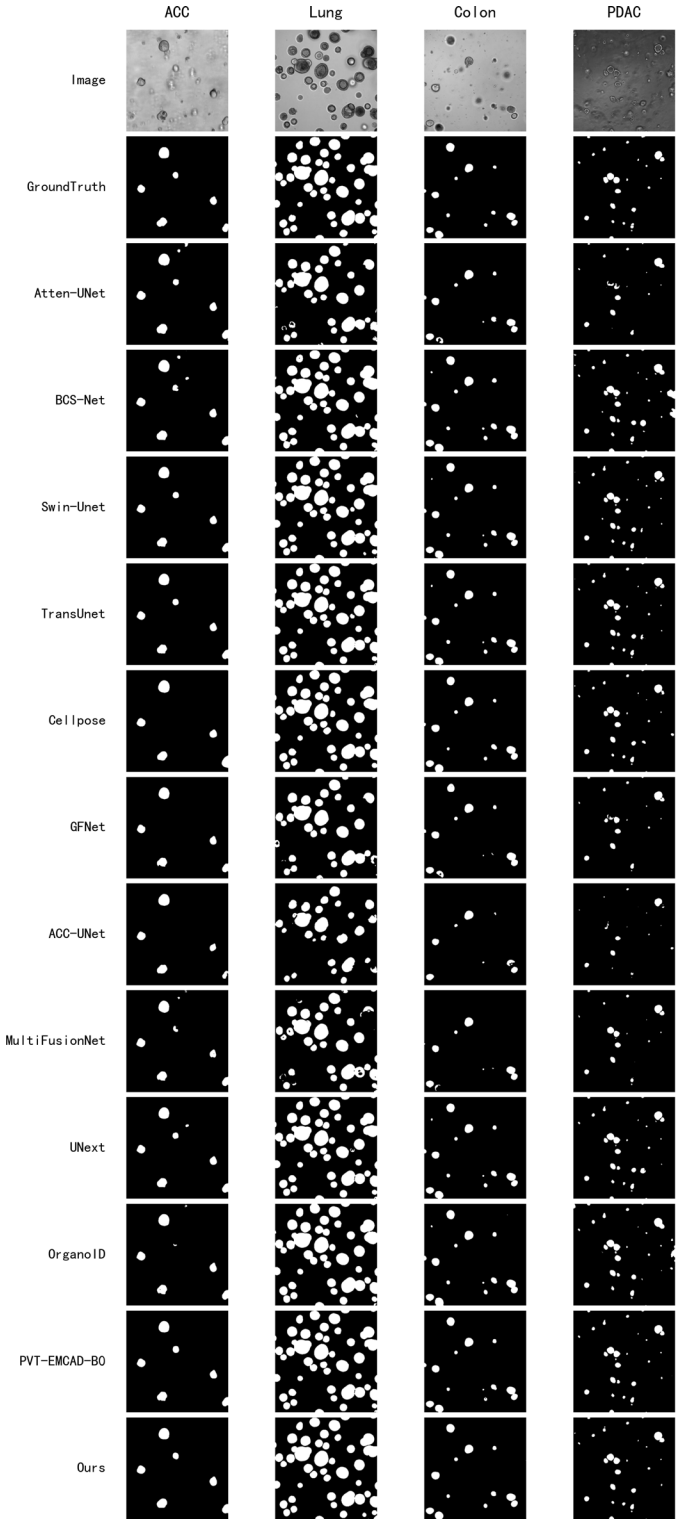

Fig. 6 Qualitative comparison of segmentation results produced by different methods on representative organoid samples from the ACC, Lung, Colon, and PDAC datasets

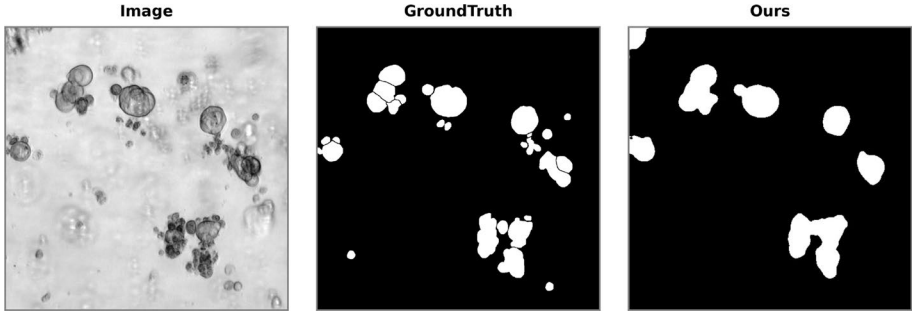

Fig. 7 Representative failure case of DeShiftNet on the ACC organoid dataset

lightweight design. Although the proposed modules enhance adaptive feature extraction, feature alignment, and multi-scale refinement, their performance in handling very weak boundaries, tightly adhered neighboring organoids, and very small fine-scale structures in challenging scenes still has room for further improvement. However, when considered together with the overall qualitative and quantitative results, DeShiftNet still shows competitive and generally more reliable segmentation performance than the compared methods while maintaining a lightweight and efficient design.

Ablation study

To quantitatively evaluate the contribution of each proposed component, we conduct an ablation study based on a self-constructed baseline that combines a standard U-Net [7] encoder with the EMCAD decoder [20]. All results are reported as averages over the merged test subsets of PDAC, ACC, Colon, and Lung organoids, ensuring that the comparison reflects performance under heterogeneous morphology and imaging conditions.

As shown in Table 3, the full model achieves the best performance, with Dice, mIoU, Precision, and Recall of 0.909, 0.836, 0.902, and 0.969, respectively. To evaluate the contribution of each proposed module, we further construct several removal-based variants from the full architecture. Removing DSE decreases the performance to 0.862/0.762/0.848/0.913, indicating that the deformable shifted encoding module is crucial for modeling organoids with heterogeneous morphology and complex boundaries. Without CAG, the model achieves 0.873/0.781/0.866/0.902, showing that cross-attention guidance enhances the interaction and alignment between encoder and decoder features. Removing D-MSCAM results in 0.880/0.792/0.871/0.893, suggesting that the proposed multi-scale context refinement improves feature aggregation and segmentation quality. Moreover, the variant without both CAG and D-MSCAM obtains 0.830/0.738/0.784/0.926, further confirming that these modules provide complementary benefits beyond DSE alone.

To further compare the proposed deformable-shift design with a simpler alternative, we replace DSE with the standard token shifting operation adopted in UNeXt [19] while keeping the remaining components unchanged. The variant with standard shift, achieves 0.877/0.798/0.875/0.924, which is inferior to the full model. Overall, the ablation results verify that DSE serves as the core module for deformable structure modeling, while CAG and D-MSCAM further enhance feature interaction and multi-scale refinement, and their joint integration yields the best overall segmentation performance.

Conclusion

In this work, we presented DeShiftNet, a lightweight and adaptive organoid segmentation framework tailored for microscopy images with heterogeneous morphology and weak boundaries. DeShiftNet integrates a deformable-shifted encoder to enhance content-aware spatial modeling, a cross-attention-guided decoder for precise encoder–decoder alignment, and a deformable multi-scale context refinement module to improve boundary coherence, all within a compact encoder–decoder design. Experimental results on the multi-type OrganoID dataset show that DeShiftNet achieves favorable segmentation performance while maintaining low model complexity, with 1.78M parameters and 2.65 GFLOPs. In addition, the model runs at 52.8 FPS with 204 MB GPU memory consumption, indicating its potential for efficient organoid image segmentation in practical applications. Ablation studies further suggest that each proposed component contributes to the overall performance improvement.

Despite these encouraging results, several limitations remain. The current evaluation is conducted on a relatively limited benchmark, and the model still has room for further improvement in challenging scenarios involving low contrast, weak boundaries, densely adhered organoids, and very small fine-scale structures. Future work will focus on enhancing boundary-aware modeling, improving the discrimination of adjacent instances and small targets, and validating the framework on larger and more diverse organoid datasets. Incorporating more explicit 3D or temporal priors may also further improve its applicability in longitudinal and high-throughput screening settings [38].

Author contributions

Conceptualization, L.T. and P.H.; Methodology, L.T., T.S. and P.H.; Investigation, T.S., X.Z. and J.B.; Resources, F.T., P.H. and L.H.; Writing - Original Draft, L.T. and T.S.; Writing - Review & Editing, L.T., T.S. and P.H.; Supervision, L.T. and P.H.; Y.H. and Z.Y. for Project Administration.

Funding

This work was supported in part by the National Natural Science Foundation of China (62302495), in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under grant (2023D01E15), in part by the Xinjiang Tianchi Talents Program (E33B9401).

Data availability

The dataset used in this study is publicly available from the OrganoID repository (https://github.com/jono-m/OrganoID) and was originally published in a previous study [21]. The source code for DeShiftNet is publicly available at: https://github.com/taost43-code/DeShiftNet.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no Competing interests.

Received: 10 December 2025 / Accepted: 16 April 2026
Published online: 27 April 2026

References

1. Corrò C, Novellasdemunt L, Li VS. A brief history of organoids. Am J Physiol Cell Physiol. 2020. https://doi.org/10.1152/ajpcell.00120.2020.
2. Waseem M, Wang B-D. Organoids: an emerging precision medicine model for prostate cancer research. Int J Mol Sci. 2024;25(2):1093.
3. Qu S, Xu R, Yi G, Li Z, Zhang H, Qi S, et al. Patient-derived organoids in human cancer: a platform for fundamental research and precision medicine. Mol Biomed. 2024;5(1):6.

4. Huang K, Li M, Li Q, Chen Z, Zhang Y, Gu Z. Image-based profiling and deep learning reveal morphological heterogeneity of colorectal cancer organoids. Comput Biol Med. 2024;173:108322.
5. Bozal SB, Sjogren G, Costa AP, Brown JS, Roberts S, Baker D, et al. Development of an automated 3d high content cell screening platform for organoid phenotyping. SLAS Discov. 2024;29(7):100182.
6. Lefferts JW, Kroes S, Smith MB, Niemöller PJ, Nieuwenhuijze ND, Kooten HN, et al. Orgasegment: deep-learning based organoid segmentation to quantify cftr dependent fluid secretion. Commun biol. 2024;7(1):319.
7. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015. pp. 234–41.
8. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. Springer; 2018. pp. 3–11.
9. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al. Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999; 2018.
10. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955; 2018.
11. Xing R. Frequ-fnet: frequency-aware u-net for imbalanced medical image segmentation. arXiv preprint arXiv:2505.17544; 2025.
12. Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, Jaeger PF. nnu-net revisited: a call for rigorous validation in 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024. pp. 488–98.
13. Su R, Zhang D, Liu J, Cheng C. Msu-net: multi-scale u-net for 2d medical image segmentation. Front Genet. 2021;12:639930.
14. Yang J, Marcus DS, Sotiras A. Dynamic u-net: adaptively calibrate features for abdominal multiorgan segmentation. In: Medical imaging 2025: computer-aided diagnosis, vol. 13407. SPIE; 2025. pp. 326–34.
15. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306; 2021.
16. Kumar S. Applications of transformer in medical imaging: a review. Impact Artif Intell Soc. 2024;1:19–30.
17. Perera S, Navard P, Yilmaz A. Segformer3d: an efficient transformer for 3d medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024. pp. 4981–8.
18. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer; 2022. pp. 205–218.
19. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. Springer; 2022. pp. 23–33.
20. Rahman MM, Munir M, Marculescu R. Emcad: efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024. pp. 11769–79.
21. Matthews JM, Schuster B, Kashaf SS, Liu P, Ben-Yishay R, Ishay-Ronen D, et al. Organoid: a versatile deep learning platform for tracking and analysis of single-organoid dynamics. PLoS Comput Biol. 2022;18(11):1010584.
22. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. Medical image segmentation review: the success of u-net. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):10076–95.
23. Lee HH, Bao S, Huo Y, Landman BA. 3d ux-net: a large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076; 2022.
24. Hofer M, Lutolf MP. Engineering organoids. Nat Rev Mater. 2021;6(5):402–20.
25. Qin Y, Li J, Chen Y, Wang Z, Huang Y-A, You Z, Hu L, Hu P, Tan F. Transorga: end-to-end multi-modal transformer-based organoid segmentation. In: International conference on intelligent computing. Springer; 2023. pp. 460–472.
26. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. pp. 4510–20.
27. Mehta S, Rastegari M, Shapiro L, Hajishirzi H. Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. pp. 9190–200.
28. Qi Y, He Y, Qi X, Zhang Y, Yang G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision; 2023. pp. 6070–9.
29. Zhang M-L, Zhang X-Y, Wang C, Liu C-L. Towards prior gap and representation gap for long-tailed recognition. Pattern Recogn. 2023;133:109012.
30. Li D, Nie X, Gong R, Lin X, Yu H. Multi-branch gan-based abnormal events detection via context learning in surveillance videos. IEEE Trans Circuits Syst Video Technol. 2023;34(5):3439–50.
31. Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ. Deep networks with stochastic depth. In: European conference on computer vision. Springer; 2016. pp. 646–61.
32. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: International workshop on deep learning in medical image analysis. Springer; 2017. pp. 240–8.
33. Cong R, Yang H, Jiang Q, Gao W, Li H, Wang C, et al. Bcs-net: boundary, context, and semantic for automatic covid-19 lung infection segmentation from ct images. IEEE Trans Instrum Meas. 2022;71:1–11.
34. Fan C, Zeng Z, Xiao L, Qu X. Gfnet: automatic segmentation of covid-19 lung infection regions using ct images based on boundary features. Pattern Recogn. 2022;132:108963.
35. Ibtehaz N, Kihara D. Acc-unet: a completely convolutional unet model for the 2020s. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023. pp. 692–702.
36. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods. 2021;18(1):100–6.
37. Agarwal S, Arya K, Meena YK. Multifusionnet: multilayer multimodal fusion of deep neural networks for chest x-ray image classification—S. Agarwal et al. Soft Comput. 2024;28(19):11535–51.

38. Bozal SB, Sjogren G, Costa AP, Brown JS, Roberts S, Baker D, et al. Development of an automated 3d high content cell screening platform for organoid phenotyping. SLAS Discov. 2024;29(7):100182.

Publisher's Note