Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076098

ARTICLE

DenT: Dense-Transformer for Label-Free Microscopy Image Segmentation

Chan-Min Hsu[1], **Shang-Ru Yang**[1], **Yi-Ju Lee**[1] **and An-Chi Wei**[1,2,*]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

*Corresponding Author: An-Chi Wei. Email: acwei86@ntu.edu.tw

Received: 14 November 2025; Accepted: 06 March 2026; Published: 08 May 2026

ABSTRACT: U-Net, a fully convolutional neural network (FCNN) with U-shaped features, has demonstrated significant success in biomedical image segmentation. However, the locality of convolution operations in the U-Net limits its ability to learn long-range dependencies. Transformers, originally developed for natural language processing, have recently been adapted for image segmentation because of their global self-attention mechanisms. Inspired by the long-range feature learning capability of transformers, we propose Dense-Transformer (DenT), an architecture designed for volumetric microscopy image segmentation. DenT incorporates transformers as encoders within each convolutional layer to capture global contextual information. Additionally, dense skip connections at multiple resolutions enhance feature propagation, enabling precise localization. We evaluated DenT on mitochondrial segmentation using our confocal microscopy dataset and a public fluorescence microscope dataset from the Allen Institute for Cell Science. The experimental results demonstrate that DenT incrementally improves the segmentation of mitochondria and mitochondrial DNA substructures from transmitted light microscopy images. DenT offers a tool for visualization, measurement, and analysis of mitochondrial morphology and mitochondrial DNA in label-free microscopy.

KEYWORDS: UNet; transformer; image segmentation; mitochondrial organelle; deep learning; microscopy imaging

1 Introduction

In recent years, deep learning has been widely adopted for image analysis applications, particularly in biomedical image segmentation [1–3]. High-performance medical image segmentation plays a crucial role in computer-aided detection and diagnosis [4,5]. Among the different deep learning-based segmentation methods, fully-convolutional neural networks (FCNNs) and their architectures [6–8] are the primary approaches. A widely used FCNN with a U-shaped structure, the so-called U-Net [7], features a symmetric encoder-decoder architecture with skip connections. The encoder extracts features from an input image using convolutional and downsampling layers, while the decoder upsamples these extracted features back to the input resolution for pixel-wise (or voxel-wise in 3D) semantic prediction. Additionally, output features of different scales from the encoder are merged into the decoder through the skip connections, allowing the recovery of spatial information lost during the downsampling process. Because of the excellent performance and ease of use, many variants of U-Net have been introduced to the area of both 2D and 3D medical segmentation [9–12], and all have achieved extensive success. In fact, many architectures, such as 3D U-Net [13], U-Net++ [14], ResUNet [11], and nnU-Net are all based on U-Net, and have been applied in various medical applications [15].

Despite their strong representation capabilities, CNN-based approaches often struggle when segmenting structures with significant shape and size variations, such as mitochondria within cells [16,17]. This limitation arises from the inherently localized nature of convolution operations, which restricts the ability of CNNs to capture global, long-range semantic information. While some studies have attempted to mitigate this issue by incorporating atrous convolutional layers [18,19], these modifications remain constrained by the fundamental locality of convolution-based processing. Consequently, CNN-based segmentation models may yield suboptimal results in some challenging datasets [20–22] where long-range dependencies are essential for accurate segmentation. On the other hand, combining a self-attention mechanism with a CNN [23,24] has been proposed to improve the capability of learning long-range semantic information.

Recently, the transformer [25], which has been highly successful in natural language processing (NLP), was introduced to the computer vision domain [26–28]. Unlike CNN-based methods, the mechanism of a transformer allows it to learn long-range dependencies. In Dosovitskiy et al. [28], a Vision Transformer (ViT) treats an image as a sequence of patch embeddings along with position embeddings that will later be used as input for image classification. TransUNet [26,29] combines a transformer with the U-Net, indicating that a transformer can be used for medical image segmentation. Both ViT and TransUNet demonstrate that a transformer can be applied to various vision domain tasks and image segmentation tasks [30,31]. Transformer-based models therefore have been adopted in computer vision and applied in medical image analysis (MIA) [32,33].

Inspired by TransUNet, we propose the Dense-Transformer (DenT), which consists of multiple transformers and dense skip connections, to leverage the power of transformers for biomedical image segmentation. To effectively learn the global context from different scales, DenT employs multiple transformers after each downsampling stage in the encoder. The extracted features from each transformer are then upsampled in the decoder. Like UNet++, we build nested and dense skip connections so that the features in the decoder can be fused with multiscale features from the encoding path for voxel-wise segmentation prediction. DenT performed as expected in extensive experiments on both our dataset of confocal microscopy images of mitochondria and DNA [17,34] and a dataset from the Allen Institute for Cell Science [12]. In this study, we position Dense-Transformer (DenT) as a task-driven architectural refinement rather than a fundamentally new segmentation paradigm. Label-free transmitted-light microscopy presents distinct challenges compared with fluorescence imaging, including low intrinsic contrast, ambiguous organelle boundaries, and densely clustered subcellular structures. Under these conditions, purely convolutional models may encounter disambiguating spatially separated but visually similar regions, whereas fully transformer-based architectures typically require large-scale pretraining, which is challenging for biological datasets. DenT is therefore designed to integrate multiscale self-attention with convolutional feature extraction in a lightweight and data-efficient manner, explicitly targeting the needs of label-free volumetric microscopy segmentation. In this context, 'label-free' refers to segmentation performed on transmitted-light microscopy images without fluorescence staining. The model is trained in a weakly supervised manner using fluorescence-derived annotations but operates on unstained transmitted-light inputs during inference.

2 Methods

2.1 Architecture Overview

In this work, inspired by TransUNet, we employ multiple CNN-Transformer blocks as encoder units to build a U-shaped architecture for 3D biomedical image segmentation, along with dense skip connections similar to the UNet++. The overall architecture of the proposed DenT is shown in Fig. 1a. Unlike existing hybrid models that apply a single transformer at the bottleneck, DenT introduces transformer blocks at multiple encoder depths. This design allows global contextual information to be captured at different spatial

resolutions, which is particularly important for resolving clustered mitochondria and fragmented structures in label-free images. Dense skip connections further stabilize optimization by promoting feature reuse and improving information flow between encoder and decoder stages, mitigating potential information loss caused by patch tokenization.

The network consists of an encoder, decoder, and nested skip connections. The basic transformer unit is adapted from the ViT and TransUNet. The goal is to predict the corresponding segmentation map from an input image size $C \times Z \times H \times W$, where Z, H, and W represent the 3D spatial resolution, and C represents the number of channels. In the encoder part, after each downsampling layer and CNN, we transform the input image into a sequence embedding by flattening the patches that are split from the image. With a patch size of $P \times P \times P$, the dimension of the sequence becomes $P \times P \times P \times C$, where the total number of patches $N = Z \times H \times W/P^3$. In addition, a linear projection is implemented on the sequence embeddings, projecting them into a D-dimensional embedding space. Similar to ViT and TransUNet, we add position embeddings to patch embeddings to preserve spatial information. The patch tokens then undergo transformer blocks to extract features in each stage. Unlike the transformer in the TransUNet, we apply several transformers in the encoder, creating multiscale transformers. In the decoder, we employ a series of CNNs that decode the high-level features back into the image with the same resolution $C \times Z \times H \times W$. To address the need for precise segmentation of high-resolution images, we implement dense skip connections [14]. Such skip pathways allow the extracted features to be fused with features from the multiscale transformers and the convolution blocks in the encoder and thus reduce the loss of spatial information during the downsampling process.

The overall architecture of the DenT is explained in the following sections.

2.1.1 Transformer Block

Transformers perform on 1D input such as words and sentences. By reshaping a 3D input into 1D sequences, we obtain the sequence $\boldsymbol{x} \in \mathrm{R}^{P^3 C}$ and number of patches $N = \frac{Z \times H \times W}{P^3}$.

In Fig. 1a,b, linear projection is applied to the patch tokens, creating a D-dimensional embedding space. The position embeddings $\boldsymbol{E}_{pos} \in \mathrm{R}^{N \times D}$ are added to embeddings $\boldsymbol{E} \in \mathrm{R}^{\left(P^3 C\right) \times D}$ as follows:

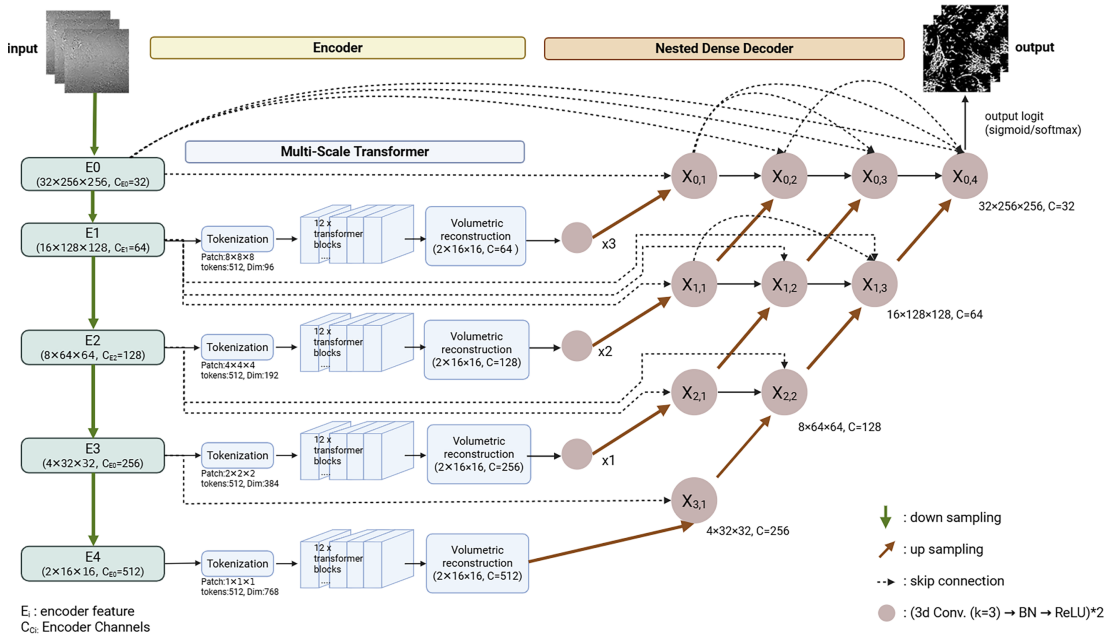
$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_p^1 \boldsymbol{E}; \boldsymbol{x}_p^2 \boldsymbol{E}; \ldots; \boldsymbol{x}_p^N \boldsymbol{E}\right] + \boldsymbol{E}_{pos} \tag{1}$$

After the embeddings, we use transformer block as encoders. The blocks include two parts: multihead self-attention (*MHSA*) and multilayer perceptron (*MLP*) or FFN (feed forward network) according to:

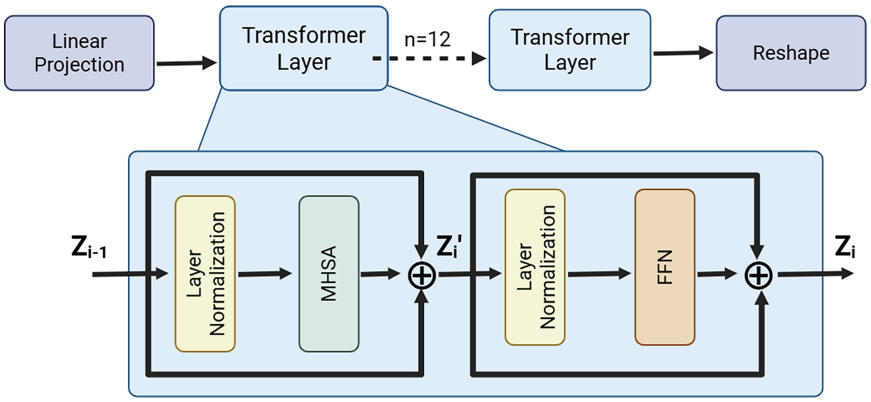
$$\boldsymbol{z}_i^{'} = MHSA\left(LayerNorm\left(\boldsymbol{z}_{i-1}\right)\right) + \boldsymbol{z}_{i-1} \tag{2}$$

$$\boldsymbol{z}_i = MLP\left(LayerNorm\left(\boldsymbol{z}_i^{'}\right)\right) + \boldsymbol{z}_i^{'} \tag{3}$$

where *LayerNorm* represents layer normalization, i is the number of layers repeated in the blocks and $\boldsymbol{z}_i$ is the input of each transformer block.



(a)



(b)

Figure 1: Architecture of DenT and transformer block details. (**a**) Overview of DenT architecture. DenT consists of a 3D convolutional encoder, a multi-scale Transformer module, and a nested dense decoder. The encoder produces multi-resolution feature maps E_0–E_4 with progressively reduced spatial dimension and increased channel widths. For scales E_1–E_4, features are tokenized using scale-specific 3D patches, yielding 512 tokens per scale with embedding dimensions. Tokens are processed by 12 Transformer blocks to capture long-range dependencies, then reshaped via volumetric reconstruction to a common low-resolution grid ($2 \times 16 \times 16$) with corresponding channels. These Transformer-enhanced features are upsampled to match the target skip resolutions and fused with encoder skip features. The nested dense decoder is organized as nodes $X_{i,j}$, where i denotes resolution level and j the nested decoding stage; each node aggregates features by dense concatenation of prior nodes in the same row and an upsampled feature from the next deeper level, followed by two $3 \times 3 \times 3$ Conv–BN–ReLU operations. The final node $X_{0,4}$ produces segmentation logits using a convolutional head, followed by sigmoid (binary) during training/inference. (**b**) Transformer block used in the multi-scale transformer. Each block applies a layer normalization, multi-head self-attention (MHSA), and residual addition, followed by layer normalization, a feed-forward network (FFN/MLP), and another residual addition. The token sequence shape is preserved throughout the transformer stack. The figures were created in BioRender.com.

2.1.2 Multiscale Transformers

In our U-shaped network, the input image first goes through CNNs and downsampling layers, similar to the U-Net. Then, multiscale transformers take the features from each convolution block as input and create 1D sequences as output. To reconstruct the 3D feature maps, we reshape the 1D sequences with size $\frac{Z \times H \times W}{P^3}$ into 3D tensors of size $\frac{Z}{P} \times \frac{H}{P} \times \frac{W}{P}$.

2.1.3 Decoder

In the decoder, the reshaped 3D feature maps are first upsampled to increase their resolution, followed by convolution layers, batch normalization (BN) layers, and rectified linear activation unit (ReLU) layers. Note that each transformer needs to perform the upsampling process multiple times (from 1 to 3), depending on the depth of the transformers.

Second, each resized feature map is concatenated with the feature maps from previous convolution blocks. After that, the concatenated part is fed into upsampling blocks, where each block consists of $3 \times 3 \times 3$ convolution layers, BN layers, and ReLU layers and a 2-times upsampling operator. This process is repeated until the input resolution is reached. In the end, the final output is passed through a $3 \times 3 \times 3$ convolution layer, creating the segmentation map.

2.1.4 Dense Skip Connection

Similar to the U-Net++ [14], we create nested, dense skip connections to better preserve spatial information. As Fig. 1 shows, not only the convolution features but also the transformer features are passed through the dense skip connections, allowing both the global and local information to be preserved. In the dense skip connection, each convolution block receives a feature map that fuses the features from the previous convolution blocks. With all the features fed into the decoder via dense skip connections, the similarity of the encoder feature map and the decoder feature map is higher than the original skip connection, which leads to better optimization [14]. The impact of the difference between a simple skip connection and a dense skip connection will be discussed in the result section.

2.1.5 Loss Function

The loss function we use is binary cross-entropy loss combined with Dice loss, which can be defined as:

$$L(X, Y) = 1 - \frac{2\sum_{n=1}^{N} X_n Y_n}{\sum_{n=1}^{N} X_n + \sum_{n=1}^{N} Y_n} - \sum_{n=1}^{N} \left(Y_n \log \sigma(X_n) + (1 - Y_n) \log (1 - \sigma(X_n)) \right) \quad (4)$$

where N is the number of voxels, and X_n and Y_n denote the predicted probability and the ground truth for voxel n, respectively.

2.1.6 Tokenization in the Multi-Scale Encoder

The model employed a multi-scale tokenization strategy to capture hierarchical representations. The tokenization process was implemented using the PatchEmbedding module. This module divided the input image into non-overlapping patches using a convolutional layer with a kernel size and stride equal to the specified patch size. The input image is partitioned into patches of varying sizes with scaling factors [2,4,8,16]. These varying patch sizes allow the model to process features at multiple resolutions, capturing both fine-grained and coarse-level spatial information.

For each scale level, feature maps are extracted using a hybrid U-Net-based feature extractor and then tokenized through convolutional layers. Each convolutional block operates with a different kernel size and stride, corresponding to the computed multi-scale patch sizes. The extracted tokens are further augmented with scale-specific positional embeddings, ensuring the model retains spatial awareness at each resolution. The four sets of positional embeddings correspond to different tokenization levels, reinforcing hierarchical feature representation. The multi-scale token sequences are processed independently before being fused within the transformer encoder.

2.2 Dataset and Evaluation

For application on high-resolution confocal microscopy images, the dataset comprises 70 sets of 3D confocal high-resolution images collected in our lab [17,34]. Imaging was performed using an LSM800 Zeiss microscope with a Plan-apochromat 1.40-NA, 63× objective, and images were acquired using Zeiss ZEN Blue 2.6 software. Three imaging channels were used: transmitted light (TL), SYBR Gold for nuclear and mitochondrial DNA labeling, and TMRM for mitochondria labeling. Each confocal image stack consists of 32 slices with an interval of 0.15 μm and a YX resolution of 917 × 917 pixels, with a pixel scale of 0.085 μm/px. To achieve uniform spatial resolution along all axes and remove voxel anisotropy, we adopted Ounkomol et al.'s method [12] to resample all z-stack volumes using cubic interpolation to obtain isotropic voxel dimensions of 0.15 μm × 0.15 μm × 0.15 μm. This resampling step removes voxel anisotropy, enabling spatially consistent learning from cubic patches (e.g., 16 × 16 × 16 voxels) in both voxel space and corresponding physical dimensions. Of the dataset, 46 samples were allocated for training, whereas 24 samples were used for validation.

For another microscopy image application, DenT was tested on the mitochondria image dataset from the Allen Institute for Cell Science [12]. The cell and organelle structures were imaged using a Zeiss spinning disk microscope with a 100×, 1.25-NA objective and Zeiss ZEN Blue 2.3 software, and mitochondria were labeled with the EGFP-tagged protein Tom20. The entire dataset was split into 60 training sets and 20 validation sets for model training. Each image consists of 50 to 75 slices with an interval of 0.29 μm, and a *YX* resolution of 624 × 924 pixels (with pixel scale of 0.108 μm px^{-1}) [12].

Both datasets were preprocessed and transformed into mask images for ground truth labeling. The preprocessing steps include histogram equalization, background subtraction, Gaussian noise removal, sigma filtering, gamma correction, and Contrast Limited Adaptive Histogram Equalization (CLAHE). Binary masks were generated by thresholding fluorescence microscopy signals (SYBR Gold and TMRM) using established intensity cutoffs. The transmitted light images used as model input were raw images without processing. Notably, these binary masks serve as the ground truth for voxel-wise binary segmentation of organelle presence. Preprocessing steps were applied only to fluorescence channels to enhance the accuracy of thresholding during ground truth mask generation. Since ground truth masks were generated by thresholding fluorescence signals and represent approximate voxel-wise semantic labels rather than instance-accurate annotations, they may introduce systematic uncertainty, particularly for small or low-contrast structures such as mitochondrial DNA. As a result, reported segmentation metrics should be interpreted as relative performance indicators rather than absolute measures of structural accuracy.

All segmentation tasks are treated as binary segmentation tasks with 1-channel input. The average Dice similarity coefficient (DSC) is used as an evaluation metric for each corresponding structure (mitochondria, DNA). The Dice score is defined as:

$$Dice\,(X, Y) = \frac{2\sum_{n=1}^{N} X_n Y_n}{\sum_{n=1}^{N} X_n + \sum_{n=1}^{N} Y_n} \tag{5}$$

where X_n and Y_n denote the value of the prediction and the ground truth for voxel n, respectively.

In addition to Dice, we also use Cross-Entropy (CE) Loss to evaluate the difference between the predicted probability distribution and the ground truth.

$$Cross\ Entropy = -\sum Y_i \log(X_i) \tag{6}$$

Y_i is the true label, and X_i is the predicted probability for each voxel.

2.3 Implementation

All the neural networks are implemented in Python 3.6 and PyTorch 1.9. The input size of an image is set to 32 × 256 × 256 for high-resolution confocal images and 64 × 192 × 128 for light microscopy images. The patch size is set to 16 × 16 × 16. We trained our models on multiple Tesla V100s with 32 GB of memory provided by the Taiwan Computing Cloud (TWCC; https://www.twcc.ai/). During the training process, the batch size is 2 unless otherwise specified. The models are trained using the Adam optimizer with a learning rate of 2e−4 and a weight decay of 5e−4. The number of training iterations is approximately 10,000 (200 epochs). We apply data augmentations such as random rotation and flipping for all the experiments. Note that to compare our model with other methods, we adjust the hyperparameters of each method individually to ensure their best performance. In the inference part, the whole image is fed into the model to reconstruct the 3D prediction.

For computational efficiency, we evaluated DenT in terms of model complexity and practical efficiency. Despite incorporating multiple transformer blocks, DenT maintains comparable parameter counts and training times (Table 1). In Table 1, computational efficiency comparison was performed on a single GPU with input size 1 × 1 × 32 × 256 × 256 and batch size = 1. FLOPs were estimated using fvcore profiling tools, representing lower-bound estimates. Empirically, training time differences across models remained within the same order of magnitude, indicating that the additional computational overhead introduced by multiscale attention is manageable in practice.

Table 1: Computational efficiency comparison of transformer-based architectures.

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)
TransUNet	132.36	404.04	102.83
DenT	162.45	1990.68	417.48

3 Results

The DenT segmentation performance on our confocal microscopy image dataset (Fig. 2), using label-free transmitted light images as input, was first evaluated by DSC, CE, and a hybrid score combining both metrics (Table 2). We also compared DenT performance with the other four methods (Table 3, Fig. 2). Given the nature of the dataset, we adapted the U-Net [7], UNet++ [14], ResUNet [11], and TransUNet [26,29] settings from 2D to 3D. The results indicate that DenT achieves the highest segmentation accuracy of 53.64% (DSC) among all the tested methods. For 3D dataset training, runtime comparisons show that DenT, U-Net, UNet++, ResUNet, and TransUNet have relatively similar training times (~4 h). Although DenT is a more complex model, its training time does not significantly exceed that of the other models, suggesting that its additional computational cost remains manageable. Compared with TransUNet, our method achieves a notable DSC improvement of 4.80%, demonstrating its effectiveness for 3D biomedical image segmentation.

Furthermore, by integrating multiscale transformers with dense skip connections, DenT enables accurate segmentation even in regions with clustered mitochondria (Fig. 2). Although the improvements achieved by DenT over CNN-based baselines are modest, they are consistent across datasets and segmentation targets. Given the inherent difficulty of label-free microscopy segmentation and the presence of annotation noise in fluorescence-derived ground truth masks, small numerical gains may still correspond to meaningful improvements in boundary delineation and structural continuity. Accordingly, DenT is not presented as a transformative performance leap, but as a stable and reproducible refinement that incrementally improves segmentation quality under challenging imaging conditions.

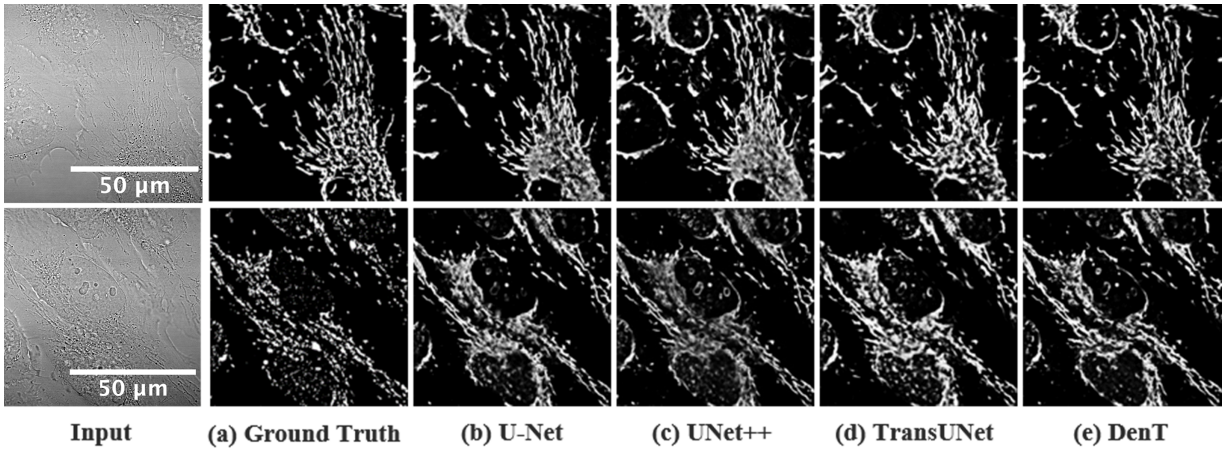


Figure 2: The segmentation results of different methods on the high-resolution confocal image dataset with transmitted light images as input [17,34]. The first row shows the mitochondrial predictions (TMRM stain) and the second row shows the mitochondrial DNA predictions (SYBR Gold stain). Scale bar: 50 µm (shown in the input images and applicable to all corresponding segmentation panels).

Table 2: The performance of the mitochondria and mitochondrial DNA segmentation task on the high-resolution confocal image dataset.

Evaluation Metric	Overall Score (%)	Mitochondria Score (%)	DNA Score (%)
Dice (DSC)	53.64	60.96	46.32
Cross Entropy (CE)	45.33	58.03	32.62
Cross Entropy + Dice (Hybrid)	53.89	61.38	46.39

Table 3: Segmentation accuracy of the different methods on the high-resolution confocal image dataset.

Model	DSC Score	Mitochondria	DNA
U-Net	52.12	58.34	45.89
U-Net++	51.56	59.36	43.75
ResUNet	48.25	55.60	40.89
TransUNet	48.84	55.61	42.06
DenT	53.64	60.96	46.32

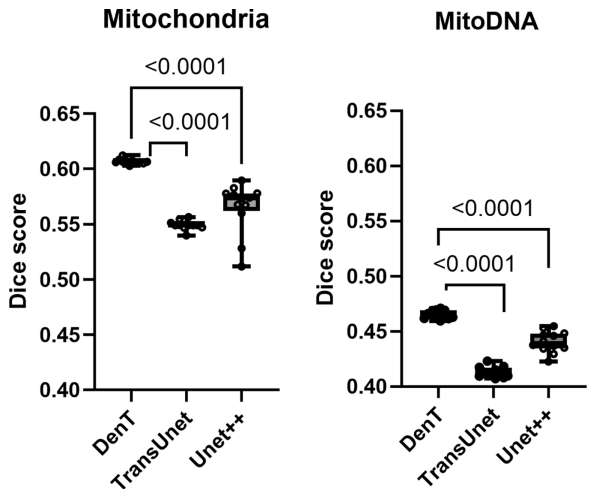


Figure 3: Performance comparison across independent runs. Dice scores of DenT, TransUNet, and UNet++ for segmenting Mitochondria (left) and MitoDNA (right) across 12 independent training runs. Each dot represents one run; the central marker and error bars indicate the mean ± SD. Brackets denote pairwise statistical comparisons between DenT and other models, with corresponding *p*-values shown above the brackets.

To evaluate DenT's applicability across different types of microscopy images, we compare its performance with that of other methods on the Allen Institute dataset (Table 4). DenT achieves good results on the mitochondrial segmentation task, demonstrating its generalizability to other datasets. Fig. 4 presents a comparison of the mitochondrial segmentation predictions on the Allen Institute dataset.

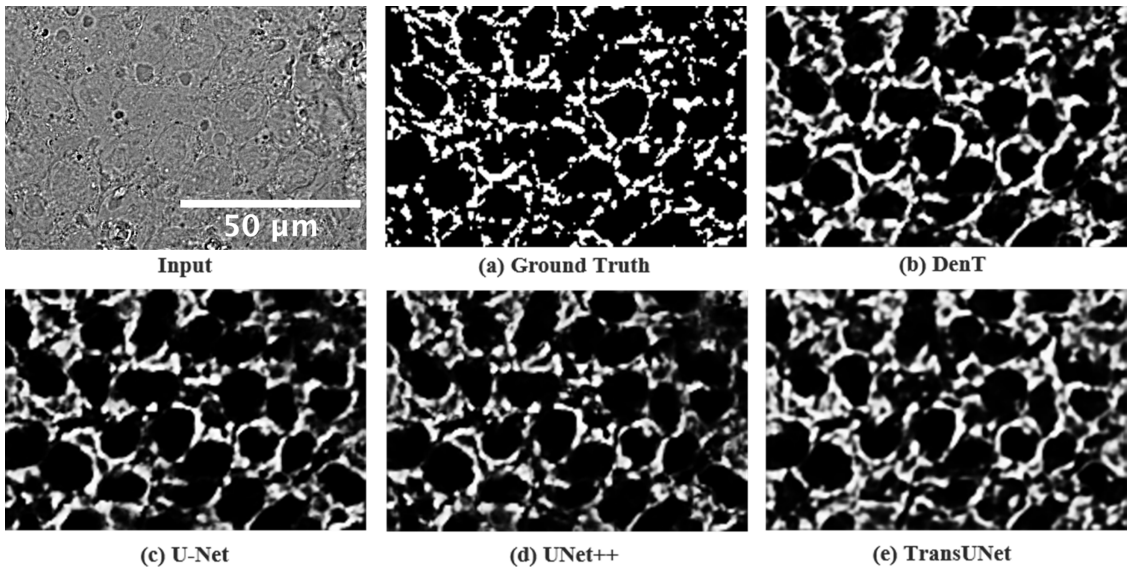


Figure 4: The segmentation results of various methods on mitochondria from the microscopy image dataset of the Allen Institute for Cell Science [12]. Scale bar: 50 μm (applies to all panels).

Table 4: Segmentation accuracy of various methods on the light microscopy image of mitochondria from the Allen Institute for Cell Science.

Model	DSC (Mitochondria)
U-Net	52.13
U-Net++	52.30
ResUNet	51.18
TransUNet	44.35
DenT	52.65

3.1 Ablation Studies

To evaluate the influence of several factors on the performance, we performed ablation studies on the dataset of high-resolution confocal images, including the decoder backbone, optimizer choice, input size, and type of skip connections.

3.1.1 Influence of the Decoder Backbone

The results of the model using different backbones in the decoder are shown in Table 5. The backbone used in our decoder is standard convolution layers followed by batch normalization and ReLU layers. Unlike TransUNet and ViT, we do not use ResNet [35] and its pretrained weights as our backbone because our dataset is 3D ($Z \times Y \times X$). Additionally, ResNet architecture typically requires large-scale datasets for effective training, making them less practical for medical imaging tasks, where annotated data are often limited. Since our dataset consists of 3D images ($Z \times Y \times X$), we opt for convolutional layers without residual connections as our backbone throughout all the experiments.

Table 5: Ablation study on the influence of the decoder backbone.

Backbone	Overall DSC Score	Mitochondria	DNA
ResUNet	49.99	57.17	42.84
Standard convolution	53.64	60.96	46.32

3.1.2 Influence of the Optimizer

Following the original settings of the TransUNet, we first used the stochastic gradient descent (SGD) method as our optimizer. However, the performance was unstable and did not yield consistently better results. Since SGD is commonly used with ResNet, we hypothesized that it might not be ideal for training DenT, which integrates transformers. Therefore, we switched to the Adam optimizer. As shown in Table 6, Adam outperformed SGD across all metrics, confirming our hypothesis regarding the choice of optimizer.

Table 6: Ablation study on the influence of the optimizer.

Optimizer	Overall DSC Score	Mitochondria	DNA
SGD	51.04	57.89	44.19
Adam	53.64	60.96	46.32

3.1.3 Influence of the Input Size

To investigate the effect of the input size, we conducted experiments on our model with different input resolutions. The results with 16 × 128 × 128 and 32 × 256 × 256 resolutions as inputs are shown in Table 7. With the same patch size of 16 × 16 × 16, the increased input size results in a larger sequence length for the transformer, consequently improving the performance of the model. The average DSC is improved by 8.98% when the input size is increased by a scale of 2. Note that a model with a lower input resolution has better computational efficiency. In our experiment, to ensure the segmentation accuracy of the model, the default resolution of the input is set to 32 × 256 × 256.

Table 7: Ablation study on input size.

Input Size	Overall DSC Score	Mitochondria	DNA
16 × 128 × 128	44.66	49.36	39.95
32 × 256 × 256	53.64	60.96	46.32

3.1.4 Influence of the Skip Connection

We further discuss the influence of different skip connections. As presented in Table 8, a dense skip connection structure leads to better performance than the simple skip connection implemented in TransUNet and others. As a result, we conclude that combining a transformer with a dense skip connection can preserve both the global and local features more effectively. Thus, we adopt the dense configuration for our model.

Table 8: Ablation study on the influence of the skip connection.

Skip Connection	DSC	Mitochondria	DNA
Simple	51.52	57.67	45.37
Dense	53.64	60.96	46.32

3.1.5 Influence of the Position Embedding

In convolutional networks, spatial relationships are inherently captured through convolution and pooling operations, which can implicitly serve as positional encodings. However, in transformer-based architecture, explicit position embeddings are often required to encode spatial information. Since DenT integrates both convolutional layers and transformers, it is important to evaluate the impact of explicit positional encoding on performance. To assess this, we conducted an ablation study on positional embeddings in DenT (Table 9).

Table 9: Ablation study on position embedding.

Position Embedding	DSC (%)	Mitochondria (%)	DNA (%)
Activate	53.89	61.3	46.39
Deactivate	53.64	60.86	46.41

The Dice Similarity Coefficient shows a small increase of 0.25% when positional encoding is activated, suggesting a limited but positive influence on overall segmentation accuracy. Additionally, positional

encoding appears to enhance the localization of mitochondrial structures more effectively than mitochondrial DNA substructures. This suggests that additional factors, such as feature representation at different resolutions, may play a more significant role in segmenting smaller substructures.

3.1.6 Influence of the Transformer Blocks

To understand the role of transformer blocks at different depths, we conducted an ablation study, sequentially removing each block and evaluating its impact on segmentation performance (Table 10). The results reveal that removing the first (shallowest) transformer block leads to the most significant drop in performance, particularly for mitochondria segmentation, while the effect of removing deeper blocks is less pronounced. This indicates that early transformer layers play a crucial role in high-resolution feature extraction, whereas later layers contribute more to global context aggregation. Since mitochondrial structures have complex shapes, early transformers enhance spatial encoding beyond the receptive field of convolutions, leading to better localization. DNA segmentation, on the other hand, may benefit slightly from later transformer blocks because mitochondrial DNA is a smaller substructure that requires context from surrounding features.

Table 10: Ablation study on transformer blocks.

Transformer Block in Different Level	DSC (%)	Mitochondria (%)	DNA (%)
DenT Model	53.89	61.38	46.39
1st Transformer Block	53.57	60.63	46.51
2nd Transformer Block	53.68	60.85	46.37
3rd Transformer Block	53.80	61.02	46.57
4th Transformer Block	54.1	61.15	47.05

4 Discussion

The development of the Dense-Transformer (DenT) model is driven by both the limitations and advancements of existing methods in medical image segmentation. The traditional U-Net, a fully convolutional neural network (FCNN) with a U-shaped structure, has been highly successful in medical image segmentation tasks due to its symmetric encoder-decoder architecture and skip connections. These design choices allow U-Net to extract and utilize multi-scale spatial information from medical images effectively. However, the reliance on convolutional operations, which are inherently local, restricts the U-Net's ability to capture long-range dependencies within the data. This limitation becomes particularly significant for tasks where understanding the global context is crucial, such as in the segmentation of complex anatomical structures or varying features across large image regions. On the other hand, transformer-based neural networks have also demonstrated success in computer vision tasks. Transformers offer an alternative to convolutional layers by enabling the model to learn long-range dependencies without being constrained by the local nature of convolutional operations. However, they often rely heavily on extensive pretraining on large-scale datasets.

Therefore, in this work, we introduce an alternative approach by using pure convolutional layers as the backbone, enabling end-to-end training of the transformer without requiring pretraining. This not only simplifies the training process but also enhances the model's adaptability to specialized datasets. DenT integrates the global self-attention mechanisms of transformers, into the U-Net architecture. By incorporating transformers as encoders at each convolutional layer of the U-Net design, the DenT aims to capture global features that are critical for accurate segmentation of biomedical images. Additionally,

the DenT introduces dense skip connections, which enhance the model's ability to fuse encoded features at different resolutions with the decoder. Dense connection methodologies inspire this design choice as in UNet++ to improve semantic information flow and facilitate localization in segmentation tasks.

Interestingly, our segmentation results for subcellular structures show that the mitochondria score consistently outperforms the mitochondrial DNA score, suggesting that the model is more effective at learning mitochondrial features. The observed performance gap between mitochondria and mitochondrial DNA segmentation likely reflects differences in annotation reliability and visual distinguishability in transmitted-light images. Mitochondria exhibit larger and more coherent morphological features, whereas mitochondrial DNA lacks chromatin organization and is more sparsely distributed, making it inherently more difficult to segment from label-free inputs. In other words, unlike nuclear DNA, which is densely packed with chromatin and forms distinguishable structural patterns, mitochondrial DNA exists in a more dispersed and less structured state within the mitochondrial matrix. This lack of chromatin-based contrast likely makes it more challenging for the model to extract clear features for mitochondrial DNA segmentation compared to nuclear DNA segmentation in our previous study [17]. Conversely, mitochondria, being larger and well-defined organelles with distinct morphological characteristics, provide more reliable spatial cues. This allows the transformer blocks to learn more effective features, contributing to the consistently higher segmentation performance of mitochondria.

While our segmentation masks do not explicitly represent mitochondrial subtypes such as puncta, rods, or networks, they provide biologically meaningful voxel-wise labels derived from validated fluorescence signals. The focus of this work is on semantic segmentation, where the objective is to identify regions of mitochondrial and mtDNA presence, rather than performing instance segmentation or morphological classification. The appearance of partial structures is a result of thresholded fluorescence signals, which is a standard and widely accepted practice in biological image annotation. Future work may incorporate shape-aware post-processing or instance-aware models to further delineate mitochondrial morphotypes. While mitochondria are considered local, in dense or low contrast regions, CNN-based models can fail to resolve boundaries between closely packed or overlapping organelles. Transformers provide global context, which helps in disambiguating such cluster-heavy regions.

We acknowledge that the size of the available training data, particularly for the high-resolution confocal dataset, is limited relative to typical transformer training requirements. This constraint may restrict the extent to which long-range dependencies can be learned and raises potential overfitting concerns. To mitigate this risk, DenT avoids large pretrained backbones and instead employs lightweight transformer blocks combined with convolutional feature extractors, dense skip connections, and extensive data augmentation. Nevertheless, future work incorporating larger datasets, self-supervised pretraining, or parameter-efficient attention mechanisms may further improve robustness and generalization.

High-resolution imaging techniques such as confocal, light-sheet, super-resolution, and electron microscopy generate rich volumetric data that require sophisticated segmentation methods [36,37]. Recent advances in microscopy and deep learning-based mitochondria segmentation have significantly improved our ability to analyze mitochondrial structures in biomedical research [38–40]. These methods have been applied to a variety of biological studies, including mitochondrial morphology analysis, disease-related mitochondrial dysfunction assessment, and cell metabolism research. Furthermore, segmentation advancements enable automated and high-throughput analysis of mitochondrial dynamics, aiding in drug discovery and personalized medicine [41–43]. The integration of advanced imaging modalities with robust segmentation algorithms continues to drive breakthroughs in mitochondrial research, enhancing our understanding of cellular bioenergetics and pathology. While DenT does not aim to redefine the state of the art in medical image segmentation, it provides a solution for label-free volumetric microscopy. By combining transformers

with the U-Net architecture, DenT leverages both convolutional and self-attention mechanisms, resulting in enhanced feature learning and improved segmentation accuracy.

5 Conclusion

In summary, DenT provides a CNN–Transformer integration for label-free microscopy segmentation. While the proposed architecture represents an incremental refinement rather than a paradigm shift, it consistently improves segmentation performance under realistic biological data constraints without relying on large-scale pretraining. DenT can provide an alternative tool for applications where fluorescence labeling is undesirable or unfavorable.

Acknowledgement: We thank the National Center for High-Performance Computing (NCHC) in Taiwan for providing computational and storage resources.

Funding Statement: This research was funded by the National Science and Technology Council in Taiwan (MOST-110-2636-B-002-017, NSTC 113-2221-E-002-048-MY3) and the National Taiwan University Center for Advanced Computing and Imaging in Biomedicine (NTU-114L900701).

Author Contributions: Conceptualization, Chan-Min Hsu and An-Chi Wei; methodology, Chan-Min Hsu; software, Chan-Min Hsu; validation, Shang-Ru Yang; formal analysis, Chan-Min Hsu and Shang-Ru Yang; data acquisition, Yi-Ju Lee; data curation, Chan-Min Hsu; writing—original draft preparation, Chan-Min Hsu and An-Chi Wei; writing—review and editing, Chan-Min Hsu, Shang-Ru Yang and An-Chi Wei. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Confocal microscopy images of AC16 cells are deposited in IEEE Dataport, doi: https://dx.doi.org/10.21227/ckc7-2t42. DenT codes are available at https://github.com/ntumitolab/DenT_.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-Net and its variants for medical image segmentation: a review of theory and applications. IEEE Access. 2021;9:82031–57. doi:10.1109/ACCESS.2021.3086020.
2. Rayed ME, Sajibul Islam SM, Niha SI, Jim JR, Kabir MM, Mridha MF. Deep learning for medical image segmentation: state-of-the-art advancements and challenges. Inform Med Unlocked. 2024;47(7):101504. doi:10.1016/j.imu.2024.101504.
3. Xia Q, Zheng H, Zou H, Luo D, Tang H, Li L, et al. A comprehensive review of deep learning for medical image segmentation. Neurocomputing. 2025;613(1):128740. doi:10.1016/j.neucom.2024.128740.
4. Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal. 2022;79:102444. doi:10.1016/j.media.2022.102444.
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42(13):60–88. doi:10.1016/j.media.2017.07.005.
6. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–51. doi:10.1109/TPAMI.2016.2572683.
7. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015; 2015 Oct 5–9; Munich, Germany. Cham, Switzerland: Springer International Publishing; 2015. p. 234–41.
8. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.

9. 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. doi:10.1109/TPAMI.2024.3435571.
10. Cai S, Tian Y, Lui H, Zeng H, Wu Y, Chen G. Dense-UNet: a novel multiphoton *in vivo* cellular image segmentation model based on a convolutional neural network. Quant Imaging Med Surg. 2020;10(6):1275–85. doi:10.21037/qims-19-1090.
11. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens. 2020;162:94–114. doi:10.1016/j.isprsjprs.2020.01.013.
12. Ounkomol C, Seshamani S, Maleckar MM, Collman F, Johnson GR. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nat Methods. 2018;15(11):917–20. doi:10.1038/s41592-018-0111-2.
13. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016; 2016 Oct 17–21; Athens, Greece. Cham, Switzerland: Springer International Publishing; 2016. p. 424–32.
14. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham, Switzerland: Springer International Publishing; 2018. p. 3–11.
15. Xiao X, Lian S, Luo Z, Li S. Weighted Res-UNet for high-quality retina vessel segmentation. In: Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME); 2018 Oct 19–21; Hangzhou, China. p. 327–31.
16. Xiao C, Chen X, Li W, Li L, Wang L, Xie Q, et al. Automatic mitochondria segmentation for EM data using a 3D supervised convolutional network. Front Neuroanat. 2018;12:92. doi:10.3389/fnana.2018.00092.
17. Hsu CM, Lee YJ, Wei AC. Convolutional neural networks predict mitochondrial structures from label-free microscopy images. In: Proceedings of the International Forum on Medical Imaging in Asia 2021; 2021 Jan 24–27; Taipei, Taiwan.
18. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48. doi:10.1109/TPAMI.2017.2699184.
19. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, et al. CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans Med Imaging. 2019;38(10):2281–92. doi:10.1109/TMI.2019.2903562.
20. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. Nat Commun. 2022;13(1):4128. doi:10.1038/s41467-022-30695-9.
21. Liu H, Li H, Wang J, Fan Y, Xu Z, Oguz I. Predicting fluorescent labels in label-free microscopy images with pix2pix and adaptive loss in Light My Cells Challenge. arXiv:2406.15716. 2024.
22. Urrea C, Vélez M. Advances in deep learning for semantic segmentation of low-contrast images: a systematic review of methods, challenges, and future directions. Sensors. 2025;25(7):2043. doi:10.3390/s25072043.
23. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207. doi:10.1016/j.media.2019.01.012.
24. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7794–803.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 6000–10.
26. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021.
27. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-Unet: unet-like pure transformer for medical image segmentation. In: Proceedings of the Computer Vision—ECCV 2022 Workshops; 2022 Oct 23–27; Tel Aviv, Israel. Cham, Switzerland: Springer Nature Switzerland; 2023. p. 205–18.

28. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
29. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal. 2024;97(2):103280. doi:10.1016/j.media.2024.103280.
30. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021; 2021 Sep 27–Oct 1; Strasbourg, France. Cham, Switzerland: Springer International Publishing; 2021. p. 36–46.
31. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: transformers for 3D medical image segmentation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA. p. 1748–58.
32. Liu Z, Lv Q, Yang Z, Li Y, Lee CH, Shen L. Recent progress in transformer-based medical image analysis. Comput Biol Med. 2023;164(16):107268. doi:10.1016/j.compbiomed.2023.107268.
33. He X, Tan EL, Bi H, Zhang X, Zhao S, Lei B. Fully transformer network for skin lesion analysis. Med Image Anal. 2022;77(1):102357. doi:10.1016/j.media.2022.102357.
34. Hsu CM, Wei AC. AC16 human cardiomyocyte cell line: SYBR gold-labeled (Thermo Fisher Scientific, Inc.) and TMRM-labeled images. Piscataway, NJ, USA: IEEE DataPort; 2021. doi:10.21227/ckc7-2t42.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
36. Gupta Y, Heintzmann R, Costa C, Jesus R, Pinho E. Deep learning-enhanced automated mitochondrial segmentation in FIB-SEM images using an entropy-weighted ensemble approach. PLoS One. 2024;19(11):e0313000. doi:10.1371/journal.pone.0313000.
37. Conrad R, Narayan K. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. Cell Syst. 2023;14(1):58–71.e5. doi:10.1016/j.cels.2022.12.006.
38. Fischer CA, Besora-Casals L, Rolland SG, Haeussler S, Singh K, Duchen M, et al. MitoSegNet: easy-to-use deep learning segmentation for analyzing mitochondrial morphology. iScience. 2020;23(10):101601. doi:10.1016/j.isci.2020.101601.
39. Ding Y, Li J, Zhang J, Li P, Bai H, Fang B, et al. Mitochondrial segmentation and function prediction in live-cell images with deep learning. Nat Commun. 2025;16(1):743. doi:10.1038/s41467-025-55825-x.
40. Michael R, Modirzadeh T, Issa TB, Jurney P. Label-free visualization and segmentation of endothelial cell mitochondria using holotomographic microscopy and U-Net. Chem Biomed Imaging. 2025;3(4):225–31. doi:10.1021/cbmi.4c00100.
41. Wang Z. Self-supervised deep learning uncovers the semantic landscape of drug-induced latent mitochondrial phenotypes. Biophys J. 2024;123(3):165a. doi:10.1016/j.bpj.2023.11.1104.
42. Chin MY, Joy DA, Samaddar M, Rana A, Chow J, Miyamoto T, et al. Novel high-content and open-source image analysis tools for profiling mitochondrial morphology in neurological cell models. SLAS Discov. 2025;31(4):100208. doi:10.1016/j.slasd.2025.100208.
43. Somani A, Sekh AA, Opstad IS, Birgisdottir ÅB, Myrmel T, Ahluwalia BS, et al. Digital staining of mitochondria in label-free live-cell microscopy. In: Bildverarbeitung für die Medizin 2021. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2021. p. 235–40.