

A Novel 3D Decoder with Weighted and Learnable Triple Attention for Segmentation of 3D Microscopy Images

Siyavash Shabani¹, Sahar A. Mohammed¹, Bahram Parvin^{†,1,2}

¹Department of Electrical and Biomedical Engineering, University of Nevada, Reno, USA

²Pennington Cancer Institute

{sshabani, bparvin}@unr.edu

Abstract

Deep neural networks are the backbone of 3D medical image segmentation architectures, showcasing exceptional application capabilities. However, their increasing model size and computational demands present significant challenges for deployment in real-world medical applications. We introduce the Weighted and Learnable Triple Attention Network (WLTA-Net), a high-performance and efficient model to further advance this area. The WLTA-Net encoder consists of a Swin Transformer, where the multi-scale outputs with different resolutions are fused by the proposed new efficient Triple Attention (WLTA) blocks at four different levels from bottom to top. To demonstrate superior performance, WLTA-Net was first evaluated on public clinical datasets, then 3D Organoid datasets with Dice and PQ scores of 93.47 ± 0.03 and 92.36 ± 0.04 , respectively. The improved performance also comes with the added benefits of reduced complexity with 35 million parameters and lower computational cost in terms of GFLOPs. The code is available here: <https://github.com/Siyavashshabani/WLTA-Net>

1. Introduction

Organoids are important biological models for drug screening and investigating biological processes[6, 7]. In cancer drug screening, the endpoint can be cell death, proliferation, or differentiation. In the latter case (e.g., differentiation), organoids need to be imaged in 3D at relatively high resolution so that colony organization and polarity markers can be profiled on a cell-by-cell basis in context [12]. Hence, an important task is the segmentation of each nucleus in a colony. Here, we propose a computational model based on multiscale triple-attention for delineating each nucleus. The proposed model is first evaluated on publicly available 3D clinical images and then applied to the in-house-generated organoid samples imaged with either confocal or deconvolution microscopy.

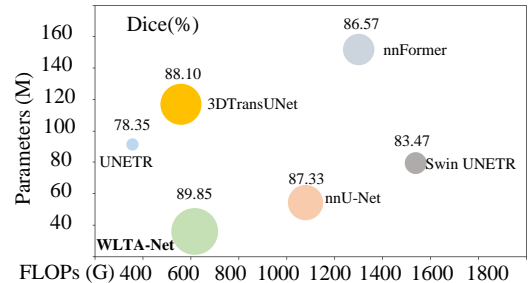


Fig. 1. WLTA-Net has the lowest number of parameters and computational loads measured by FLOPs on the Synapse dataset.

Vision Transformers (ViT [16]) and convolution [18] modules have been used extensively as the backbone in many vision applications, including but not limited to 3D medical image segmentation [20]. Among these, Swin UNETR [8, 23] leverages the Swin Transformer to construct a U-Net-shaped architecture, incorporating a Transformer-based encoder and a convolutional-based decoder. Although these models have increased quantitative and qualitative results, their performance, number of parameters, and computational costs are open questions and restrict them in real-world applications such as 3D organoid image segmentation.

One of the most critical factors influencing the performance of a deep learning model is its architectural design. Developing an efficient structure that achieves high performance remains a long-standing challenge in 3D Medical Image Segmentation. The most popular end-to-end models follow the UNet [24] Shape-based structure, which contains encoder and decoder blocks with skip connections. Attention-based models[8, 19] also adhere to this structure, with the primary distinction of utilizing ViTs and their derivatives as encoders. Recently, UNet++ [25], DiNTS [26], cascade decoder [27], and CoTr [28] have introduced more complex architectures (e.g., dense skip connection and attention mechanism) and optimization techniques for cross-scale feature fusion.

[†]Corresponding author.

This study introduces a decoder framework based on the Weighted and Learnable Triple Network (WLTA-Net), which replaces direct skip connections between encoder and decoder with Weighted and Learnable Triple Attention (WLTA) modules, which results in the lower number of parameters and improved computational loads, as shown in Figure 1. The model architecture consists of four layers of WLTA, as shown in Figure 2. In Layer 0, each WLTA module integrates the embedded outputs of Swin Transformers at three consecutive resolutions. At each consecutive layer, the outputs of the WLTA from previous layers are integrated in the same fashion. We suggest that this new merging mechanism has a direct effect on creating more contextual high-level fused features between the encoder and decoder. The efficacy of the new complement decoder layout was first validated on two public datasets of clinical images of Synapse² and ACDC³ and then applied to an in-house 3D organoid dataset.

The main contributions of our study are: (i) Designing a simple embedding module that transfers high-level feature maps of different resolutions into a unified embedded space, (ii) proposing a novel efficient Triple Attention (WLTA) block to merge embedded features from three different resolutions effectively, and (iii) demonstrating the efficiency of WLTA-Net with 35M and lower GFLOPs.

The organization of this paper is as follows: Section 2 provides a brief review of prior research in context. Section 3 outlines the methodology and its implementation. Section 4 summarizes this study's results, visualization, ablation study, and conclusion.

2. Related works

This section summarizes relevant studies for segmenting clinical images and cellular structures imaged with MRI and 3D microscopy.

2.1. Evaluation with Clinical Images

In this section, we summarize the recent studies in the field of medical image segmentation. More advanced studies [29, 30] have focused on variants of Transformer-based architectures to enhance the performance of medical segmentation tasks. For example, these enhancements are made possible in (i) 3D TransUNet [14] by integrating the Transformer blocks in the bottleneck (e.g., the bottom of a U-shaped model); (ii) nnFormer [19] by integrating local self-attention modules in the encoder and decoder modules and global self-attention in the bottleneck; (iii) Swin UNETR [8] by integrating a Swin transformer and a CNN-based model for encoder and decoder blocks, respectively; (iv) UNETR++ [21] by reducing the quadratic time

complexity of the self-attention module to a linear one; and (v) MAT3D [31] by incorporating the Multi-Aperture mechanism [32, 33] of four Swin transformers in the encoder blocks coupled with a custom-designed distance transform loss; (vi) segformer3d [17] by encoding feature maps at different scales of the input volume following the pyramid of ViT and using the shallow MLP layers as a decoder.

2.2. Evaluation with Microscopy Images

In organoid models [34-36], primary challenges in 3D nuclear segmentation stem from (i) the lack of high-resolution microscopy in the Z-direction and (ii) the inability to thoroughly wash the fluorescent dye from the matrix scaffolding. The latter contributes to (a) a halo effect around an organoid and (b) clumping of adjacent nuclei. Traditionally, researchers have employed classical image analysis methods to resolve the clumping of nuclei, which have been extensively reviewed [37]. However, deep learning (DL) methods generalize better for clumped nuclei and the diversity of cellular phenotypes [38]. The emergence of 2D [39] and 3D U-Nets [24] and its derivatives [3, 40] marks a revolutionary step, enabling deep learning to surpass traditional computer vision techniques in 3D nuclei segmentation.

The U-Net model is based on a convolutional neural network with skip connections between the encoder and decoder and has been successfully extended and applied to 3D nuclei segmentation with customized loss functions that better partition clumps of nuclei [41]. Although U-Net can capture local information, its ability to model long-range dependencies is hindered. Hence, some researchers [42] have incorporated components (e.g., attention gating) in the decoder block to improve the performance of 3D nuclei segmentation. On the other hand, Li et. al. [43] proposes a U-Net-style architecture in which the encoder block comprises Swin Transformer modules followed by convolutional modules in the decoder block. Han et al. [44] propose a hybrid framework consisting of two parallel networks, Mask R-CNN [45] and Swin Transformer, whose outputs are merged through a specialized fusion block at the end of both networks.

In summary, while previous studies have combined attention-based and convolutional blocks to develop high-performance models, the challenge of designing an efficient decoder that effectively propagates long-range contextual dependencies for prediction remains an open research question. This study addresses this gap by proposing an efficient attention-based decoder.

² <https://www.synapse.org/#!Synapse:syn3193805/wiki/217789>

³ <https://www.creatis.insa-lyon.fr/Challenge/acdc/>

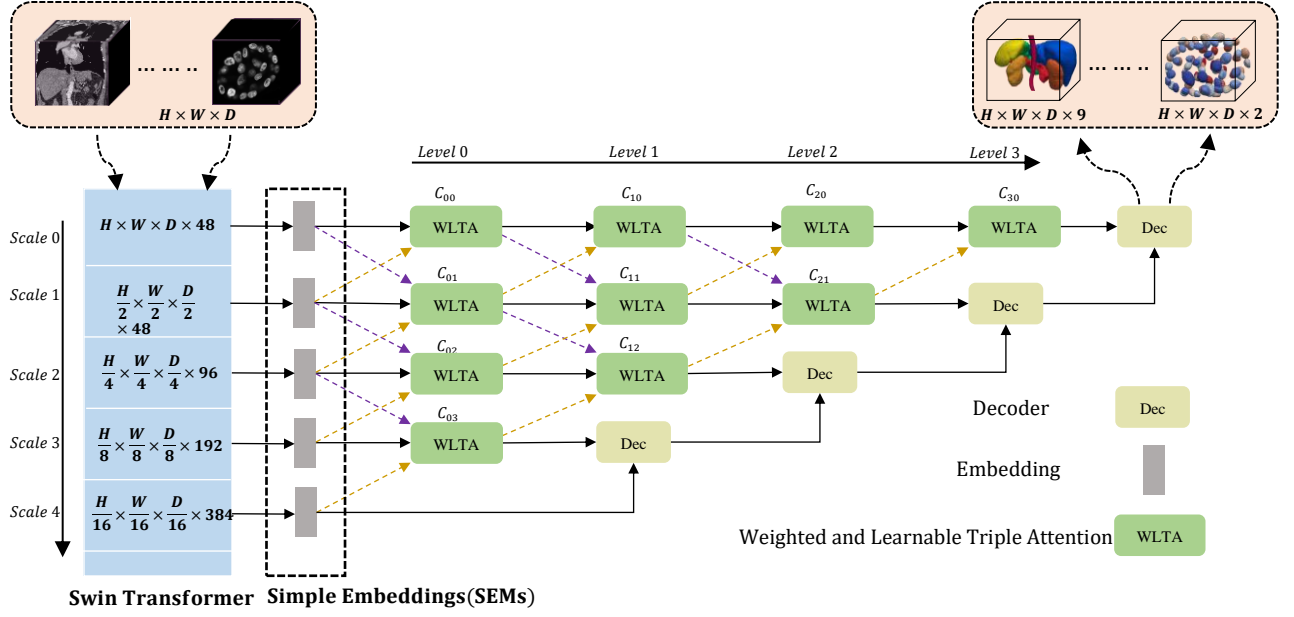


Fig. 2. WLTA-Net maps hierarchical high-level features extracted by the encoder (Swin Transformer) into an embedding space. These features, combined in pairs or triplets at different resolutions, are progressively merged through WLTA modules across four sequential levels. The refined features are then processed by shallow deconvolutional blocks to reconstruct the predicted mask.

3. Method

This section describes WLTA-Net, its submodules, and its implementation. Figure 2 illustrates WLTA-Net, where the Swin Transformer backbone extracts hierarchical high-level features in five multi-scale resolutions. These features are then mapped to the same dimension in embedded space through simple embedding modules, followed by WLTA modules and decoder blocks to reconstruct the predicted mask. The following sections describe WLTA-Net and its constituent blocks in more detail.

3.1. The Architecture

The proposed architecture is shown in Figure 2. Given the 3D patch input image $I_{in} \in \mathbb{R}^{D \times H \times W}$, the Swin Transformer backbone extracts the feature maps at different resolution scales, represented by $x = (x^{i_0, c_0}, x^{i_1, c_1}, \dots, x^{i_n, c_n})$, where c_n are the feature map sizes ($c_0, c_1, c_2, c_3, c_4, c_5$) = (48, 48, 96, 192, 384) at each scale and i is the decreasing scale (1, 2, 4, 8, and 16) of feature sizes. The feature map at each scale i is a tensor with dimensions $\frac{h}{i_n} \times \frac{w}{i_n} \times \frac{d}{i_n} \times c_n$, where h, w, d and c are

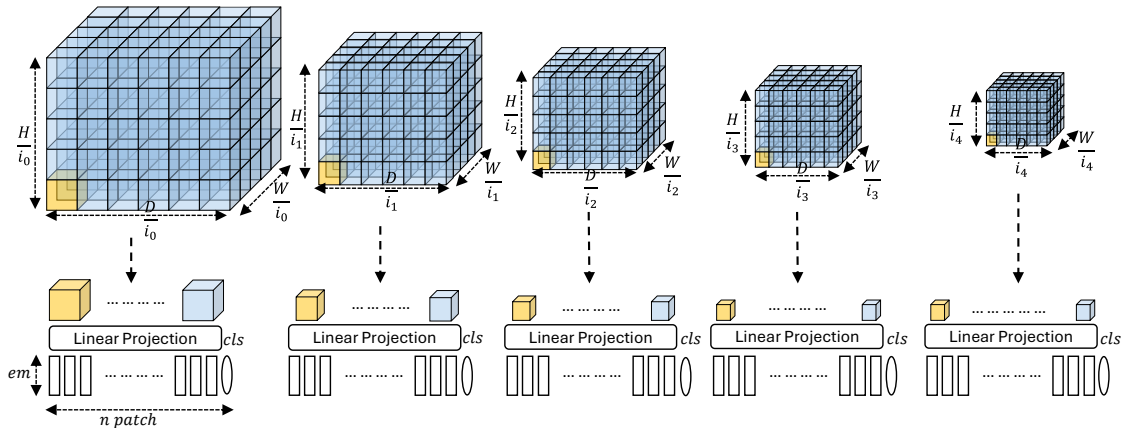


Fig. 3. SEMs project 3D hierarchical high-level features at five different resolutions into a unified embedding space.

height, width, depth, and the number of channels at the scale i , respectively.

The hierarchical feature maps of the Swin Transformer are projected with Simple Embedding into a lower-dimensional space. Then they are aggregated through multiple stages with WLTA modules. As depicted in Figure 2, at level i ($0 < i \leq L - 1$) of this process, the features at the scale j in the current stage are the result of fusing the adjacent features from the previous level along three directions: 1. Down-flow (in dotted red) high-resolution feature $x^{i-1,j-1}$ which contains high-level feature maps, 2. Forward-flow (in black) maintains the spatial resolution $x^{i-1,j}$, 3. Up-flow (in dotted blue) high-resolution feature $x^{i-1,j+1}$ which contains low-level feature maps. The fused feature maps at the i -th level of the j -th scale can be formulated as:

$$x^{i,j} = \begin{cases} WLTA(x^{i,0}, x^{i,1}), & j = 0 \\ WLTA(x^{i-1,j-1}, x^{i,j}, x^{i+1,j+1}), & j \geq 1 \end{cases} \quad (1)$$

The details of the Simple Embedding module are presented below.

3.2. Simple Embedding Module (SEM)

Figure 3 illustrates the SEM at each of the five resolutions, i.e., $(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \text{ and } \frac{1}{16})$. The embedding space is a triplet that consists of the batch size (B), number of patches (n_patch), and embedding size (em). Each linear projection layer consists of a convolutional layer with the same kernel size, stride, and *patch_size* as hyperparameters. The hyperparameter tuning rendered a 16-by-16 kernel size. The training included batch normalization and dropout.

3.3. The WLTA Module

Figure 4 shows the integration of the three inputs and their corresponding self-attention modules. Let i be the resolution level and:

$$\begin{aligned} s_{q_i} &= em_i \cdot w_{q_i}, \\ s_{k_i} &= em_i \cdot w_{k_i}, \\ s_{v_i} &= em_i \cdot w_{v_i} \end{aligned} \quad (2)$$

And for the second input ($i + 1$):

$$\begin{aligned} s_{q_{i+1}} &= em_{i+1} \cdot w_{q_{i+1}}, \\ s_{k_{i+1}} &= em_{i+1} \cdot w_{k_{i+1}}, \\ s_{v_{i+1}} &= em_{i+1} \cdot w_{v_{i+1}} \end{aligned} \quad (3)$$

And for the third input ($i + 2$):

$$\begin{aligned} s_{q_{i+2}} &= em_{i+2} \cdot w_{q_{i+2}}, \\ s_{k_{i+2}} &= em_{i+2} \cdot w_{k_{i+2}}, \\ s_{v_{i+2}} &= em_{i+2} \cdot w_{v_{i+2}} \end{aligned} \quad (4)$$

Where the em_i, em_{i+1}, em_{i+2} denotes as the inputs of different scales, and w_q, w_k, w_v are learnable weights of query, key, and value, respectively.

Next by inspiring by self-attention [16], we calculate the attention for each input by:

$$\begin{aligned} att_i &= s_{q_i} \cdot s_{k_i}^T, \\ att_{i+1} &= s_{q_{i+1}} \cdot s_{k_{i+1}}^T, \\ att_{i+2} &= s_{q_{i+2}} \cdot s_{k_{i+2}}^T, \end{aligned} \quad (5)$$

$$\begin{aligned} att'_i &= \text{softmax}(att_i) \cdot s_{v_i}, \\ att'_{i+1} &= \text{softmax}(att_{i+1}) \cdot s_{v_{i+1}}, \\ att'_{i+2} &= \text{softmax}(att_{i+2}) \cdot s_{v_{i+2}} \end{aligned} \quad (6)$$

Then, we combined a trainable weighted aggregating equation to sum up the outputs of attention modules:

$$S_w = w_i \cdot att'_i + w_{i+1} \cdot att'_{i+1} + w_{i+2} \cdot att'_{i+2} \quad (7)$$

Where learning these weights leads to adaptive attention modules. In the final step, the norm and fully connected layers are designed to get the final outputs:

$$\text{output} = \text{MLP}(\text{LN}(S_w)) \quad (8)$$

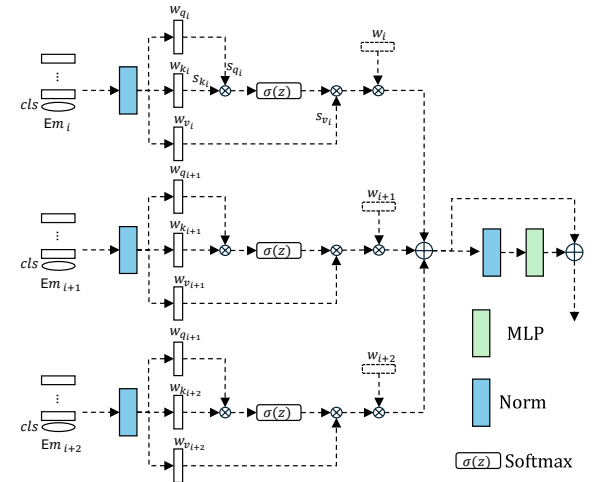


Fig.4. The Weighted Learnable Triple Attention (WLTA) module merges triple-embedded features using learnable weights for adaptive feature fusion.

3.4. Datasets

In this study, for the evaluation of WLTA-Net, we used two public datasets: Synapse multi-organ segmentation and ACDC datasets, and a private dataset for 3D organoid segmentation.

Synapse multi-organ segmentation dataset comprises 30 abdominal CT scans from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, totaling 3779 axial contrast-enhanced abdominal clinical CT images. Each CT volume consists of 85 ~ 198 slices of $[512 \times 512]$

pixels, with a voxel spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])mm^3$. We followed [8] protocol to report the average Dice and average Hausdorff Distance (HD) [46] on eight abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, and stomach) with a random split of 18 training cases, 6 cases for validation, and 6 cases for testing.

The ACDC dataset includes 100 3D cardiac MRI samples, per MIST [47] manuscript, collected from different subjects, each with a corresponding segmentation mask. Each MRI scan contains three compartments: the right ventricle (RV), the myocardium (Myo), and the left ventricle (LV). The training, validation, and testing ratios are at 70, 20, and 10 samples.

The 3D Organoid dataset encompasses organoid cultures fixed on various days, representing a diverse mutation landscape of breast cancer cell lines, including MCF10A, MCF7, MDA-MB-231, and MDA-MB-468. MCF10A cells are non-malignant, while MCF7 cells are estrogen (ER) and progesterone receptor (PR) positive but ERBB2 negative. In contrast, MDA-MB-231 and MDA-MB-468 are triple-negative breast cancer (TNBC) cell lines, lacking ER, PR, and ERBB2 receptors. In 3D cultures, these cell lines exhibit various structures such as hollow spheres, solid spheres, sheet-like formations, or grape-like clusters [34]. Despite both being TNBC, MDA-MB-231 and MDA-MB-468 display distinct phenotypes. This diversity is crucial for our research as we aim to classify primary cell phenotypes based on their 3D structure. This dataset was imaged with either a confocal or a wide-field microscope with apotome and deconvolution software. The pixel dimensions are maintained at 0.25, 0.25, and 1 micron in X, Y, and Z,

respectively. The cameras for the confocal and the wide-field microscopes have a dynamic range of 14 and 16 bits, respectively. All images are normalized between 0 and 1 and reshaped into an isotropic space for processing.

3.5. Implementation Details

Experiments were conducted on a Linux Cluster Server equipped with 8 NVIDIA RTX 3080 GPUs, each with 12 GB of memory. We followed the method described in [8]; all images of the Synapse and ACDC datasets were resized with the same physical dimensions of $1 \times 1 \times 1$ mm; subsequently, $96 \times 96 \times 96$, random patches were extracted to build a dataset with the augmentation techniques outlined in [8]. The model was trained for 300 epochs for patches randomly selected from training images. Next, testing was performed on test images, where the best-performing model within the 300 epochs was saved on the validation dataset during the training process. The final performance is reported using 5-fold cross-validation based on Adam optimization with a learning rate of 10^{-4} , and a weight decay of 5×10^{-5} .

3.6. Evaluation Metrics

To evaluate segmentation performance across three datasets, we use three metrics: the Dice coefficient (Dice), Hausdorff distance (HD95), and Panoptic Quality (PQ) [48]. The Dice score measures pixel-level consistencies, and PQ and HD95 scores measure object-level consistencies, where the PQ score is defined as follows:

$$PQ = \frac{|TP|}{|TP| + 0.5|FP| + 0.5|FN|} \times \frac{\sum_{(y,\hat{y}) \in TP} IoU(y,\hat{y})}{|TP|} \quad (9)$$

Methods		Avg		Spl	Kid(R)	Kid(L)	Gal	Liv	Sto	Aor	Pan
		Dice \uparrow	HD95 \downarrow								
nnU-Net [3]		87.33	-	91.68	88.46	84.68	78.82	<u>97.13</u>	83.34	93.04	81.50
SAMed [4]		81.88	20.64	88.72	79.95	80.45	69.11	94.80	82.06	87.77	72.17
U-Mamba [7, 9]		82.83	27.17	92.33	85.83	86.31	63.98	96.22	79.18	90.55	68.27
MIST [13]		86.92	11.07	92.83	<u>93.28</u>	92.54	74.58	94.94	87.23	89.15	72.43
Swin UNETR [8]		83.47	10.55	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80
nnFormer [19]		86.57	10.63	90.51	86.25	86.57	70.17	96.84	86.83	92.04	<u>83.35</u>
UNETR++ [21]		87.22	07.53	<u>95.77</u>	87.18	87.54	71.25	96.42	<u>86.01</u>	92.52	81.10
MISSFormer [22]		81.96	18.20	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67
LeVit-UNet [11]		78.53	16.84	88.86	80.25	84.61	62.23	93.11	72.76	87.33	59.07
3DTransUNet [14]*		88.10	-	93.39	87.47	85.76	81.15	97.34	85.31	<u>92.97</u>	81.76
WLTA-Net	LDice	89.04	13.56	94.65	92.56	92.51	78.12	93.34	83.5	92.71	82.45
	LFocal	89.17	12.10	94.41	92.69	<u>92.73</u>	79.41	93.81	83.71	91.49	82.68
	LCeDice	89.85	<u>10.90</u>	95.89	93.49	93.20	<u>79.94</u>	94.40	84.69	93.21	84.39

Table 1. WLTA-Net architecture improves the average dice score on the synapse dataset over the prior art. The upper part of the Table lists the performance of the SOTA models. The last three rows of the lower part of the Table show the performance of WLTA-Net with the three most used loss functions (Dice, Focal, and CeDice). Evaluation is performed against multiple organs (Gal for Gallbladder, Kid (L) for Left Kidney, Kid (R) for Right Kidney, Pan for Pancreas, Spl for Spleen, Sto for Stomach, Liv for Liver, and Aor for Aorta) and reported as the mean dice scores. The highest and second-highest scores are in bold and underlined, respectively. The results are reported following five-fold cross-validation. *The HD95 for 3DTransunet has not been reported in their original manuscript.

Here, TP represents True Positives, the correctly identified nuclei; FP stands for False Positives, which are the nuclei incorrectly identified as being present; and FN denotes False Negatives, which are the nuclei that are present but were not detected. y and \hat{y} correspond to the ground truth and prediction, respectively. Finally, $\text{IoU}(y, \hat{y})$ denotes the intersection-over-union.

4. Results

WLTA-Net was first evaluated on publicly available clinical 3D Synapse and ACDC datasets and then against the state-of-the-art techniques for 3D organoid segmentation.

WLTA-Net was first evaluated on publicly available clinical 3D Synapse and ACDC datasets and then against the state-of-the-art techniques for 3D organoid segmentation.

4.1. Results of Synapse Dataset

Table 1 provides a comprehensive performance profile for WLTA-Net and the contribution of each loss term as compared to the SOTA models with five-fold cross-validation. WLTA-Net records a mean Dice score of

Methods		Dice↑	RV	Myo	LV
VIT-CUP [2]		81.45	81.46	70.71	92.18
R50-VIT-CUP [2]		87.57	86.07	81.88	94.75
Swin UNETR [8]		88.61	85.29	86.52	94.02
LeViT-UNet [11]		90.32	89.55	87.64	93.76
SAM3D[15]		90.41	89.44	94.67	87.12
SegFormer3D [17]		90.96	88.50	88.86	95.53
nnFormer [19]		92.06	90.94	89.58	95.65
U-Mamba [9]		92.22	91.83	90.22	94.54
MIST [13]		92.56	91.23	90.31	96.14
UNETR++ [21]		92.83	<u>91.89</u>	<u>90.61</u>	<u>96.00</u>
WLTA-Net	L _{Dice}	91.94	91.2	89.23	95.39
	L _{Focal}	91.30	90.5	88.59	94.81
	L _{CeDice}	93.20	92.48	90.6	96.53

Table 2. WLTA-Net has the best on the ACDC dataset. The upper part of the Table lists the performance of the SOTA. The last three rows of the Table list the performance of WLTA-Net with L_{Dice}, L_{Focal}, and L_{CeDice} loss functions.

89.85 \pm 0.04 and HD95 10.90 \pm 0.03 with five-fold cross-validation. In addition, out of eight classes of organs, we observed improvement in five: Spleen (spl), Right Kidney (Kid(R)), Left Kidney (Kid(L)), Aorta (Aor), and Pancreas (Pan). Figure 5 provides a comparative visualization of the 3D segmentation results of four image samples with Swin UNETR and 3D TransUNet. This result illustrates reduced spurious regions and continuity within each organ, which are visible to Swin UNETR.

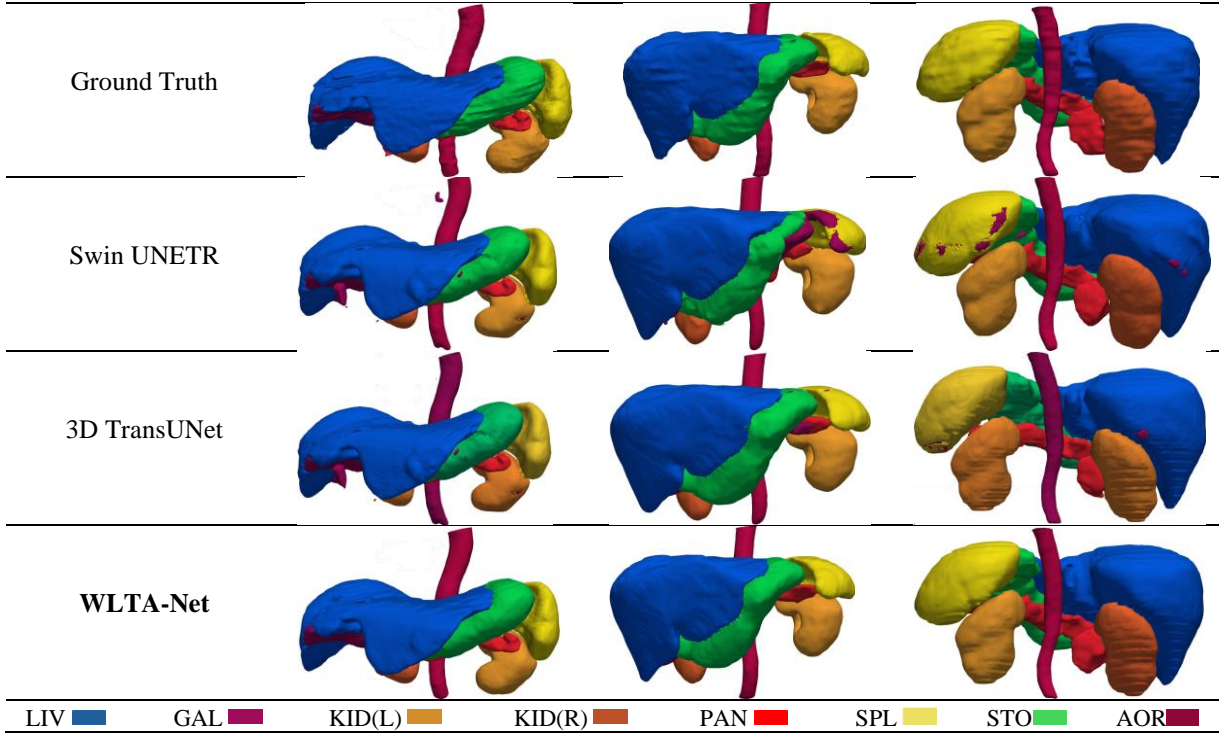


Fig. 5. WLTA-Net shows an improved qualitative behavior when compared to the prior art. Three independent samples, one per column, are shown. The results indicate that WLTA-Net has (i) less spurious noise compared to Swin UNETR, and 3D TransUNet, (ii) little or no fragmentation compared to other models, (iii) fewer artifacts in GAL compared with 3D TransUNet.

4.2. Results of ACDC Dataset

Table 2 compares the performance of WLTA-Net and different loss functions against SOTA architectures, as the upper section of the Table lists the prior research. The performance of WLTA-Net with different loss functions is presented in the lower section. Accordingly, WLTA-Net with proposed L_{CeDice} loss achieves an average Dice score of 93.20 ± 0.05 , which is better than other methods and performs better on two out of three organs—the right ventricle (RV) and left ventricle (LV). Figure 6 compares Swin UNETR, nnFormer, and WLTA-Net qualitatively. WLTA-Net is closer to the ground truth and has less fragmentation.

4.3. Results of 3D Organoid Dataset

We compared WLTA-Net against last year's top-performing method for the organoid dataset. Table 3 presents a summary of the evaluation findings. WLTA-Net, employing three different loss functions (as shown in the lower three rows), demonstrates superior performance compared to Swin UNETR [8] and other methods [1, 5, 14, 39]. It achieves a high Dice score of 93.47 ± 0.03 and a PQ score of 92.36 ± 0.04 . This comparison underscores the advancements in segmentation techniques. Figure 7 presents segmentation results and compares them with other SOTA models.

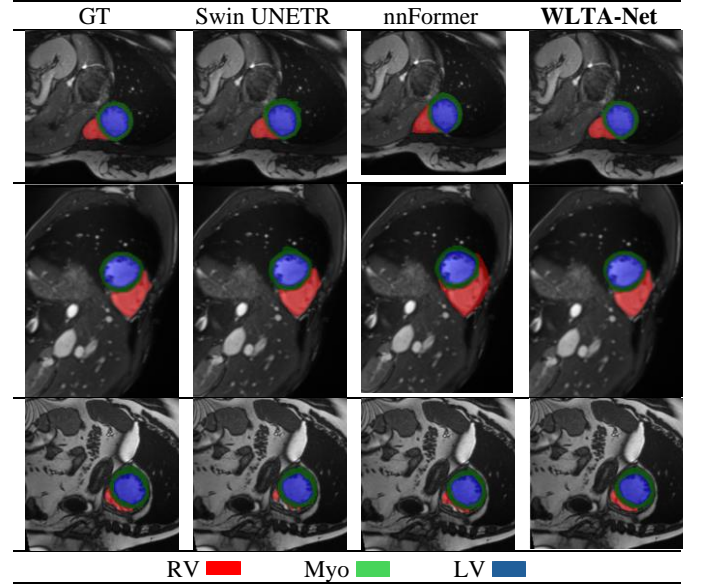


Fig.6. WLTA-Net compares better qualitatively with the ground truth without fragmentation. This is observable in the right ventricle (RV) shown in red. Comparison may require zooming.

5. Ablation Study

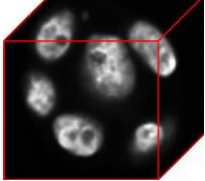
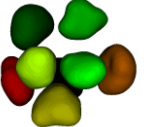


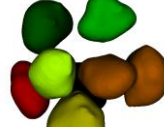
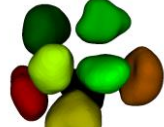
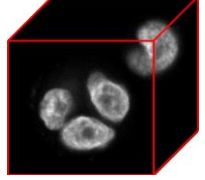
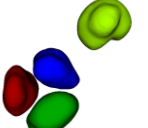
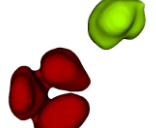
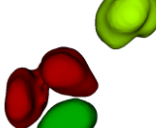
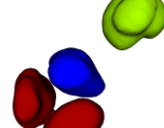
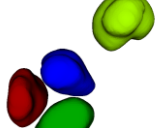
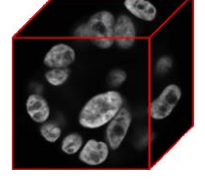

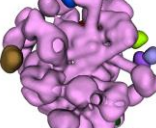
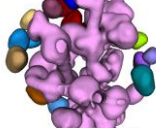
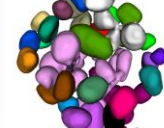
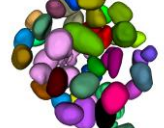
3D Organoid	Ground Truth	3D U-Net	3DTransUNet	Swin UNETR	WLTA-Net
					
No.of cells	8	6	6	7	8
					
No.of cells	4	2	3	3	4
					
No.of cells	46	13	18	31	41

Fig.7. WLTA-Net shows superior segmentation quality compared to prior methods, ensuring that adjacent nuclei are accurately delineated without merging. For example, (i) 3D U-Net significantly merges adjacent nuclei; (ii) the number of predicted nuclei better correlates with the number of nuclei in the ground truth (GT) using our model.

5.1. Effect of Backbone Architecture

As shown in Figure 2, we utilized the Swin Transformer as the default backbone of WLTA-Net and 3D U-Net [24] as an alternative backbone. Table 4 presents a performance comparison, demonstrating that WLTA-Net with the Swin Transformer outperforms the 3D U-Net backbone. Specifically, for the Synapse dataset, the average Dice Score improves by approximately 2.5% when using the Swin Transformer. Furthermore, the Swin Transformer backbone exhibits lower FLOPs and fewer parameters, highlighting its computational efficiency.

Method		Dice	PQ	Pars
OTSU Thresholding [1]		64.67	59.76	N/A
SAM 2D [5]		77.73	70.26	91M
3D U-Net [7, 10]		80.65	79.34	95M
3DTransUNet [14]		86.46	85.09	118M
Swin UNETR [8]		89.74	86.00	62M
WLTA-Net	L _{Dice}	92.26	89.56	
	L _{Focal}	91.54	90.92	35M
	L _{CeDice}	93.47	92.36	

Table 3. WLTA-Net improves quantitative indices for the segmentation of nuclei on organoids.

Backbone	GFLOPs↓	Pars↓	Dice↑	HD95↓
3D UNet	720	44	87.25	14.90
Swin Transformer	630	35	89.85	10.90

Table 4. Ablation study of the effectiveness of Swin Transformer and 3D UNet as the backbone of WLTA-Net.

5.2. Behavior of the WLTA Module

The WLTA module has three embedded inputs (em_i , em_{i+1} , and em_{i+2}), which are merged after passing through their corresponding attention modules with learnable averaging weights (w_i , w_{i+1} , and w_{i+2}). Table 5 presents the effect of alternative merging mechanisms based on max and constant weights, i.e., (0.25, 0.5, 0.25). The results indicate the superior performance of learnable weights on different datasets.

To further investigate the behavior WLTA module, we examined the learnable weights of the upper and lower

Merge	Weights of WLTA			Synapse		ACDC
	em_i	em_{i+1}	em_{i+2}	Dice↑	HD95↓	Dice↑
Max	-	-	-	86.81	15.90	88.47
Sum	1	1	1	86.19	14.94	90.74
	0.25	0.50	0.25	87.14	13.84	91.57
	0.50	0.25	0.25	87.26	13.78	91.42
	0.25	0.25	0.50	87.19	14.92	90.84
	L	L	L	89.85	10.90	93.20

Table 5. Ablation study on merging mechanism of WLTA module for Synapse and ACDC datasets. L: Learnable.

branches of the $WLTA_{01}$ block during training. As shown in Figure 8, our examination indicates that at the beginning of the training process, the model emphasizes low-level features (e.g., edges) by increasing the weight of the upper branch. However, as training progresses, particularly during the middle and final stages, the model gradually shifts its focus toward high-level features (e.g., composition) by increasing the weight of the lower branch accordingly.

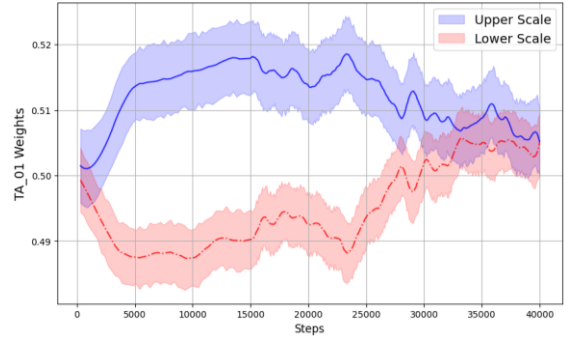


Fig.8. Variation of learnable weights in the upper and lower branches of the $WLTA_{01}$ block during training with five-fold validation.

6. Conclusion

We proposed a new framework, which includes Swin Transformers as the backbone and a new decoder module based on weighted and learnable triple attention for 3D segmentation of organoid samples imaged in 3D. The enhanced performance of the proposed framework was first evaluated on publicly available 3D clinical datasets and then on a dataset of organoids that were imaged in 3D using confocal or deconvolution. By maintaining high-resolution details and effectively integrating the outputs of the Swin Transformer via WLTA modules, improved performance is demonstrated while reducing the number of parameters to 35M. We showed an improved Dice score of 89.85 and 93.20 for Synapse and ACDC datasets, respectively. For segmenting nuclei in the organoid datasets, enhanced performance is shown in the Dice and PQ scores of 93.47 and 92.36, respectively. Moreover, the number of cells in an organoid is biologically significant because it relates to proliferation under control and treatment conditions. The proposed method provides a more consistent count of the cells between the ground truth and predicted ones. These advancements motivate a broader application of WLTA-Net in clinical imaging and drug screening with organoid models.

Acknowledgment

This research is supported by a grant from NIH RO1-CA279408.

1. References

- [1] X. Xu, S. Xu, L. Jin, and E. Song, "Characteristic analysis of Otsu threshold and its applications," *Pattern recognition letters*, vol. 32, no. 7, pp. 956-961, 2011.
- [2] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [3] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203-211, 2021.
- [4] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.
- [5] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015-4026.
- [6] C. C. Bilgin, G. Fontenay, Q. Cheng, H. Chang, J. Han, and B. Parvin, "BioSig3D: high content screening of three-dimensional cell culture models," *PloS one*, vol. 11, no. 3, p. e0148379, 2016.
- [7] V. Srivastava, T. R. Huycke, K. T. Phong, and Z. J. Gartner, "Organoid models for mammary gland dynamics and breast cancer," *Current opinion in cell biology*, vol. 66, pp. 51-58, 2020.
- [8] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20730-20740.
- [9] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III* 18, 2015: Springer, pp. 234-241.
- [11] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make faster encoders with transformer for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2023: Springer, pp. 42-53.
- [12] G. Y. Lee, P. A. Kenny, E. H. Lee, and M. J. Bissell, "Three-dimensional culture models of normal and malignant breast epithelial cells," *Nature methods*, vol. 4, no. 4, pp. 359-365, 2007.
- [13] M. M. Rahman, *et al.*, "MIST: Medical Image Segmentation Transformer with Convolutional Attention Mixing (CAM) Decoder," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.
- [14] J. Chen *et al.*, "3d transunet: Advancing medical image segmentation through vision transformers," *arXiv preprint arXiv:2310.07781*, 2023.
- [15] N.-T. Bui *et al.*, "Sam3d: Segment anything model in volumetric medical images," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024: IEEE, pp. 1-4.
- [16] A. Dosovitskiy, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale.," *arXiv preprint*, 2020.
- [17] S. Perera, P. Navard, and A. Yilmaz, "SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4981-4988.
- [18] S. Niyas, S. Pawan, M. A. Kumar, and J. Rajan, "Medical image segmentation with 3D convolutional neural networks: A survey," *Neurocomputing*, vol. 493, pp. 397-413, 2022.
- [19] H.-Y. Z. e. al., "nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer," *IEEE Transactions on Image Processing*, pp. 4036-4045, 2023.
- [20] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, pp. 1-28, 2015.
- [21] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "UNETR++: delving into efficient and accurate 3D medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.
- [22] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "MISSFormer: an effective transformer for 2D medical image segmentation," *IEEE transactions on medical imaging*, 2022.
- [23] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574-584.
- [24] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II* 19, 2016: Springer, pp. 424-432.
- [25] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 2018: Springer, pp. 3-11.
- [26] Y. He, D. Yang, H. Roth, C. Zhao, and D. Xu, "Dints: Differentiable neural network topology search for 3d medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5841-5850.
- [27] P. Liang, J. Chen, H. Zheng, L. Yang, Y. Zhang, and D. Z. Chen, "Cascade decoder: A universal decoding method for biomedical image segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019: IEEE, pp. 339-342.

- [28] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III* 24, 2021: Springer, pp. 171-180.
- [29] S. J. e. al., "MetaSeg: Content-Aware Meta-Net for Omni-Supervised Semantic Segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [30] E. Xie, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, 2021.
- [31] M. Sohaib, S. Shabani, S. A. Mohammed, G. Winkelmaier, and B. Parvin, "Multi-Aperture Transformers for 3D (MAT3D) Segmentation of Clinical and Microscopic Images," presented at the Proceedings of the Winter Conference on Applications of Computer Vision (WACV), 2025.
- [32] S. Shabani, M. Sohaib, S. A. Mohammed, and B. Parvin, "Multi-Aperture Fusion of Transformer-Convolutional Network (MFTC-Net) for 3D Medical Image Segmentation and Visualization," *arXiv preprint arXiv:2406.17080*, 2024.
- [33] S. Shabani, S. Mohammed, M. Sohaib, and B. PARVIN, "Maca-Net: Multi-Aperture Curvature Aware Network for Instance-Nuclei Segmentation," *Available at SSRN 5146469*.
- [34] J. Han *et al.*, "Molecular predictors of 3D morphogenesis by breast cancer cell lines in 3D culture," *PLoS computational biology*, vol. 6, no. 2, p. e1000684, 2010.
- [35] M. Sohaib, S. Shabani, S. A. Mohammed, G. Winkelmaier, Q. Cheng, and B. Parvin, "3D-Organoid-SwinNet: High-content profiling of 3D organoids through Transformer-Based Architecture," in *IEEE-EMBS International Conference on Biomedical and Health Informatics*.
- [36] M. Sohaib, S. Shabani, S. A. Mohammed, and B. Parvin, "3D-Organoid-SwinNet: High-content profiling of 3D organoids," *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [37] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review," *IEEE reviews in biomedical engineering*, vol. 9, pp. 234-263, 2016.
- [38] F. Kromp *et al.*, "Evaluation of deep learning architectures for complex immunofluorescence nuclear image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1934-1949, 2021.
- [39] O. Ronneberger, Philipp Fischer, and Thomas Brox., "U-net: Convolutional networks for biomedical image segmentation," *Medical image computing and computer-assisted intervention–MICCAI*, 2015.
- [40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016: Ieee, pp. 565-571.
- [41] G. Winkelmaier and B. Parvin, "An enhanced loss function simplifies the deep learning model for characterizing the 3D organoid models," *Bioinformatics*, vol. 37, no. 18, pp. 3084-3085, 2021.
- [42] L. Wu, A. Chen, P. Salama, S. Winfree, K. W. Dunn, and E. J. Delp, "Nisnet3d: Three-dimensional nuclear synthesis and instance segmentation for fluorescence microscopy images," *Scientific Reports*, vol. 13, no. 1, p. 9533, 2023.
- [43] Z. Li, Z. Huang, Z. Zhu, S. You, Z. Zhao, and H. Yan, "Directional And Topological Transformer With Topology Priors For 4D Cellular Image Segmentation," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024: IEEE, pp. 2902-2908.
- [44] Y. Han *et al.*, "An ensemble method with edge awareness for abnormally shaped nuclei segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4315-4325.
- [45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.
- [46] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proceedings of 12th international conference on pattern recognition*, 1994, vol. 1: IEEE, pp. 566-568.
- [47] M. M. Rahman, S. Shokouhmand, S. Bhatt, and M. Faezipour, "MIST: Medical Image Segmentation Transformer with Convolutional Attention Mixing (CAM) Decoder," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 404-413.
- [48] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404-9413.