Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5, 3293-3308
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i5.7679

Hybrid U-net based on channel reconstruction and self-attention calibration bridge for weakly-supervised cell segmentation

Shuping Yuan[1,2*], Vladimir Y. Mariano[1]
[1]College of Computing and Information Technologies, National University, Manila, Philippines; y_shu_ping@163.com (S.Y.) vymariano@national-u.edu.ph (V.Y.M.).
[2]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, China.

Abstract: Semantic segmentation of cellular images obtained via optical microscopy is a critical area in medical image analysis. However, the irregularity of cellular images on glass slides, along with excessive impurities, presents significant challenges for existing segmentation algorithms, leading to frequent semantic misclassification. This paper introduces a Hybrid U-Net architecture based on encoding channel reconstruction and self-attention calibration bridging to achieve weakly supervised extraction of cellular category information. The architecture, built upon a hybrid CNN and transformer module framework, first proposes channel reconstruction during the convolutional encoding phase, creating a latent space for channel information by expanding the dimensions. Subsequently, texture and global information within the latent space are compressed and reconstructed, effectively eliminating inter-channel redundancy at minimal additional cost. Additionally, a self-attention calibration bridging module is employed in the skip connections, where attention is extracted from shallow features in two dimensions. The proposed method is tested and validated on the publicly available SIPaKMeD dataset, outperforming other semantic segmentation networks in terms of mIoU and Dice scores.

Keywords: *Channel reconstruction, Hybrid U-Net architecture, Optical microscopy, SIPaKMeD dataset, Weakly-supervised learning.*

1. Introduction

Semantic segmentation of cellular images [1] obtained through optical microscopy is a core technology in biomedical image analysis, aimed at accurately separating cellular regions from image backgrounds or other tissue structures. This task plays a critical role in fields such as cell counting [2] pathological diagnosis [3] drug screening [4] and disease research [5]. With the continuous advancement of microscopy imaging technologies, the quality of cellular images has significantly improved, providing more detailed information for the segmentation task. However, cellular images often exhibit complex morphological characteristics, such as significant variations in cell shape, size, and density, as well as issues like adhesion, overlapping, and blurred boundaries between cells, making semantic segmentation of cells increasingly challenging.

Early cell segmentation methods primarily relied on classical image processing techniques such as thresholding [6] edge detection [7, 8] region growing [9, 10] and morphological operations [11]. While these methods could achieve certain results in simpler cases, they typically perform poorly in complex cellular images due to their limited robustness to noise, morphological variations, and cell overlap. Additionally, traditional methods struggle to fully utilize high-level semantic information from the images, limiting improvements in segmentation accuracy.

With the rapid development of computer vision and deep learning, Convolutional Neural Networks (CNNs) [12] have made breakthrough advancements in the field of image segmentation. Deep learning methods, by automatically learning prior knowledge from large amounts of labeled data, effectively

History: Received: 13 March 2025; Revised: 6 May 2025; Accepted: 9 May 2025; Published: 30 May 2025
* Correspondence: y_shu_ping@163.com

overcome the limitations of traditional methods, significantly improving segmentation accuracy. Notably, deep learning models based on architectures such as Fully Convolutional Networks (FCN), U-Net [13, 14] and Mask R-CNN [15, 16] are capable of capturing multi-level features of cell morphology and structure, using contextual information from the images for accurate segmentation. These models can not only recognize individual cells but also handle cell cluster segmentation [17, 18] in complex scenarios.

Currently, research in cell semantic segmentation is continuously expanding and deepening. Recent studies have focused on multi-scale feature fusion algorithms [19] as an important approach to enhance segmentation performance. By extracting features at different scales, these methods effectively address the challenges posed by differences in cell size and morphology. Moreover, incorporating various attention mechanisms [20] allows models to focus on the most important regions of the image, thereby avoiding background noise interference and improving accuracy. In weakly-supervised [21] cell segmentation tasks, image augmentation techniques help achieve good segmentation results even with limited data, and the application of semi-supervised [22] learning and transfer learning [23] has greatly advanced the research in cell semantic segmentation.

U-Net models have been widely applied to various medical image segmentation tasks with significant progress, but challenges remain in semantic segmentation tasks for cell clusters with incomplete labels. First, although U-Net architecture exhibits excellent fitting ability and robustness on small-scale datasets, it requires a large amount of high-quality manual annotations as prior knowledge to effectively train the correct weights. The complexity and diversity of cell cluster images make the annotation process time-consuming and heavily reliant on expert knowledge, and it is difficult to accurately annotate all cell shapes. Second, although U-Net architecture with skip connections can effectively address cell overlap and occlusion, segmentation errors still occur when cells are highly overlapped or boundaries are blurred.

In this paper, based on the hybrid U-Net architecture incorporating CNN and transformer units, we optimize the features in both the encoding and skip connection stages and apply them to weakly-supervised cell cluster segmentation tasks. Our method enhances the initial encoding stage's understanding of channel interactions through the introduction of a CNN-based channel reconstruction module. We also use a self-attention calibration bridging module in the skip connection stage to decouple and reconstruct features from three different scales along different axes, generating region-specific attention. The main contributions of this work are summarized as follows:

i. We reconstructed the training set for the cell cluster segmentation dataset with incomplete labels, converting the weakly-supervised segmentation problem into a strongly-supervised task, yielding good performance during inference.
ii. We built a feature channel reconstruction module at the backend of the encoding stage across different scales, enhancing the inter-channel interactions and eliminating redundant information in the channel dimensions.
iii. We constructed a self-attention calibration bridging module for the skip connection stage, enabling self-attention to be reapplied to the initial features during the transmission process, effectively eliminating background interference and various types of noise in cell cluster segmentation tasks.

2. Related Work

2.1. Cell Segmentation

Researchers have developed numerous algorithms to address the challenges of cell recognition and segmentation, which mainly rely on traditional image processing techniques and deep learning-based computer vision algorithms. Traditional image processing techniques typically extract specific features and perform threshold segmentation and morphological classification based on characteristics such as texture, color, shape, and entropy of the cells. For instance, Veta, et al. [24] employed predefined nuclear geometries and watershed algorithms for cluster nuclei separation, while Wienert, et al. [25]

utilized morphological operations without involving watershed. Liao, et al. [26] used eclipse fitting methods for cluster separation. These traditional methods rely heavily on specialized prior knowledge and require pre-designed tuning of hyperparameters, which, in practice, have limitations. They are often inadequate for handling the complex background interference in cell segmentation tasks and struggle to correctly identify features of cells in various biological tissues.

On the other hand, with the widespread application and development of deep learning in image segmentation, many neural network-based end-to-end segmentation algorithms have also been applied to cell image segmentation tasks. Li, et al. [27] proposed a segmentation method based on white blood cell localization and deep semantic U-Net, achieving semantic segmentation on a small white blood cell dataset. Fan, et al. [28] introduced a "white blood cell membrane" system that includes localization and segmentation branches. Tran, et al. [29] designed a fully automated white blood cell and red blood cell counting system based on SegNet, but neglected the segmentation of nuclear components. Additionally, Mahbod, et al. [30] adopted a hybrid model combining U-Net and attention mechanisms, such as SSD and U-Net, to perform instance segmentation of cells within clusters.

2.2. Weakly-Supervised Semantic Segmentation

Weakly-supervised semantic segmentation (WSSS) algorithms primarily generate pseudo-labels to create attention-based class activation maps [31, 32] from which segmentation masks are reconstructed. Past research has led to heuristic advancements in WSSS, such as adversarial erasure, aimed at guiding the network to explore new regions rather than being limited to salient areas. Additionally, techniques such as self-supervised learning [31, 33] contrastive learning [34] and cross-image information [35] have been employed to enhance the accuracy and completeness of CAMs. Recently, visual language pre-training techniques [36] have emerged as an effective approach for addressing visual-language tasks, with notable applications in WSSS [37]. On the other hand, due to the coarse boundaries of initial attention maps, refinement methods such as CRF [38] and IRN [39] have been used. However, segmentation results generated from activation maps face issues like misalignment and blurred boundaries compared to fully-supervised methods. In this work, we propose a data reconstruction approach to address the problem of insufficient labels for cell cluster segmentation tasks.

3. Method

Given a batch of incomplete cell samples of size $N \times \mathbb{R}^{H \times W \times 3}$, where N denotes the number of samples, and R represents the dimensions of the cell cluster's confocal microscopy imaging with $H \times W \times 3$, we first reconstruct the incomplete labeled dataset by removing the unlabeled portions and reformulating it into a fully labeled dataset. This process transforms the weakly supervised task into a strongly supervised one, with specific details provided in section 3.2. Next, an end-to-end multi-class semantic segmentation is performed using a hybrid-Unet architecture. Finally, we validate the results on an additional manually labeled test set. This section introduces our approach through data reconstruction, the channel reconstruction module during the encoding phase, and the self-attention calibration bridging module.

DOI: 10.55214/25768484.v9i5.7679

3.1. Overall Architecture

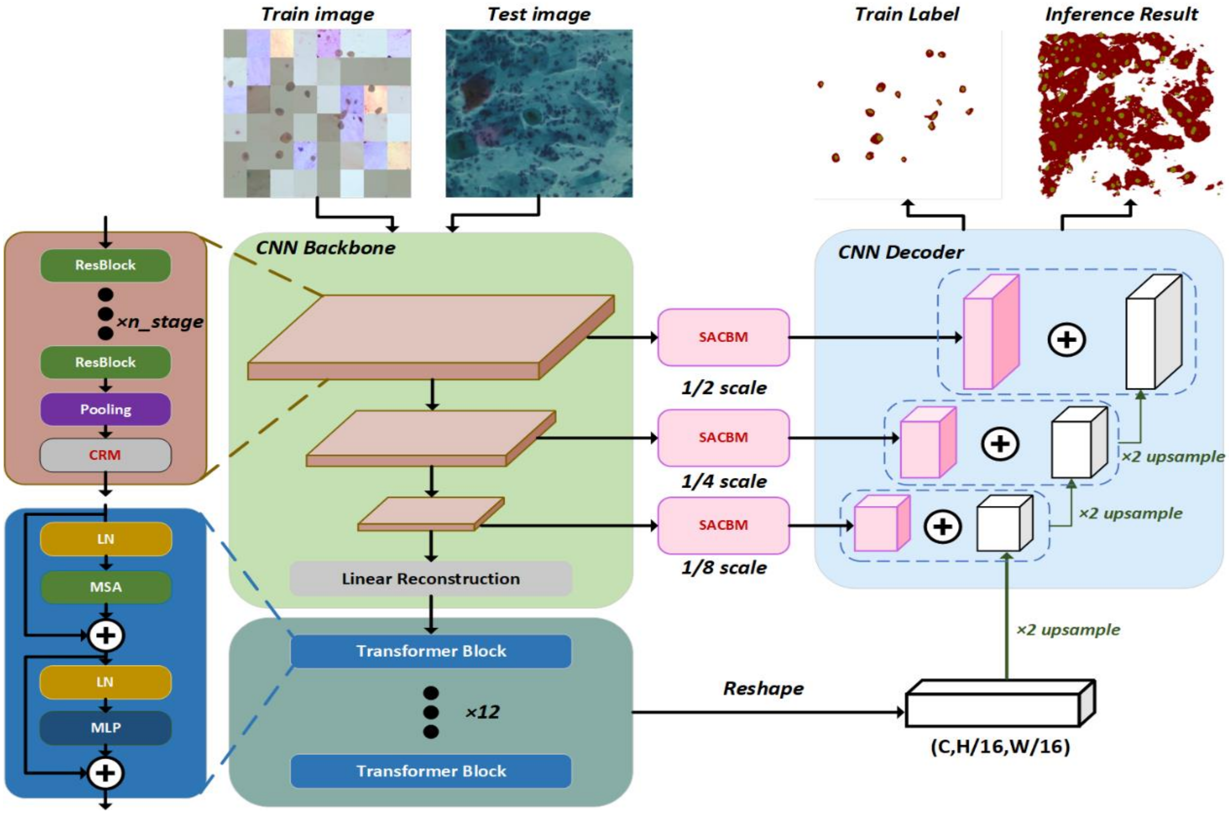

Figure 1.
Weakly Supervised Hybrid U-Net Segmentation Flowchart.

The weakly supervised hybrid U-Net segmentation framework constructed in this study is shown in Figure 1. The encoder structure consists of three residual convolution blocks at different scales and twelve layers of transformer blocks. The CNN encoder progressively reduces the input image scale to 1/2, 1/4, and 1/8, with each CNN unit containing several residual blocks, a pooling layer, and a channel reconstruction layer to represent 2D spatial features. The output features are then converted into a linear sequence and input into the transformer encoder. After passing through twelve transformer blocks with multi-head self-attention (MSA) mechanisms, the features are restored to 2D representations. Subsequently, four stages of upsampling are performed, with the corresponding scale feature maps from the CNN backbone decoupled and refocused via the self-attention calibration bridging module during the skip connection phase of each upsampling stage. These are then concatenated, followed by upsampling and classification to obtain the test labels and inference results. The training dataset shown in Figure 1 consists of cell cluster images reconstructed from raw samples, with the generation process described in section 3.2. The test input is the raw sample.

DOI: 10.55214/25768484.v9i5.7679

3.2. Weakly Supervised Data Reconstruction

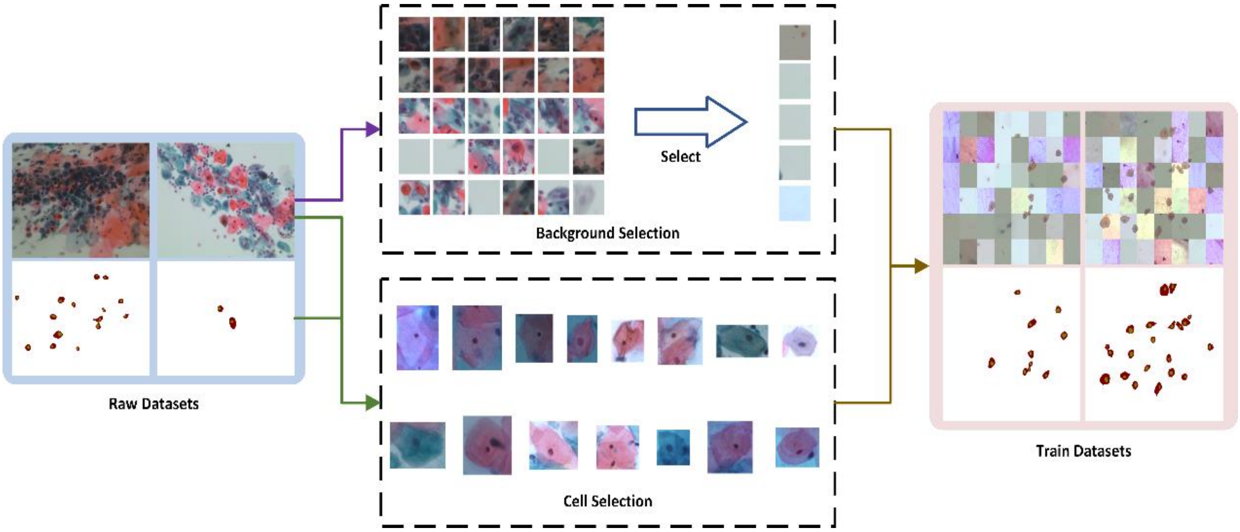

Figure 2.
Weakly Supervised Data Reconstruction Process

There are three main challenges in image processing and semantic segmentation of cell cluster data:

The boundaries of cells are unclear, with each individual cell being distinct and exhibiting significant variability in shape, color, and features. Cell boundaries are often accompanied by issues such as blurring and fading, which can cause confusion with the background or other cells.

Overlapping regions between individual cells are prevalent. In multi-cell images, cells are not typically isolated; cells of the same type tend to overlap and connect, making it difficult to distinguish which cell the nucleus and cytoplasm belong to, as well as causing intersection and union problems due to cytoplasm overlap.

The sampling background contains impurities that cause significant interference. For example, some impurities appear as black circular shapes, which are difficult to distinguish from the cell nucleus in terms of morphology and color, leading to missegmentation and misclassification.

Therefore, as shown on the left side of Figure 2, manual labeling of each cell is a significant challenge. Relying solely on visual observation and experience can only allow for labeling a very small number of cells in each image, resulting in incomplete samples. Incomplete data labeling negatively impacts the performance and test results of weakly supervised semantic segmentation tasks in two main ways: First, in terms of quantity, a multi-cell image typically contains several to dozens of cells. Due to the aforementioned reasons, most cells cannot be labeled, leading to a small proportion of correctly labeled single-cell images, generally not exceeding 20%. Second, in terms of dataset size, due to the limited data collection and the low proportion of positive samples in the entire sample space, overfitting and long-tail effects are likely to occur during network training, making it difficult to obtain correct weight parameters for testing.

To address these issues, we have developed a data augmentation method suitable for cell cluster data to expand and reconstruct the dataset. Unlike traditional methods such as morphological reconstruction, adding noise, and random rotation or translation, we use a random concatenation approach for dataset reconstruction. The specific process is as follows: As shown in Figure 2, we name the entire process "qualitative selection concatenation," which consists of two main workflows:

Background selection. We split each cell cluster image into 6x8 sections, resulting in 48 background blocks. After secondary screening of these background blocks to retain those without cells, we randomly select and reorganize these blocks to form the background of the new dataset.

Cell selection. We extract labeled single-cell images from the cell cluster. The extraction process involves selecting a bounding box centered around the cell nucleus's position in the image, using the average single-cell size. These bounding boxes are then filtered to retain only the areas containing the current cell, ensuring that the selected box is not disturbed by other cells or impurities, forming negative samples.

After these two steps, the extracted single-cell images are multiplied with their corresponding labels, and the single-cell results are randomly rotated, translated, and scaled, then placed into the background generated in the first step to create the final augmented dataset. Although the generated samples may appear disjointed and incomplete to the human eye, using native backgrounds in computer vision significantly enhances the neural network's denoising, recognition, and robustness capabilities, helping the network better fit realistic cell cluster segmentation scenarios.。

3.3. Channel Reconstruction in the Encoding Stage

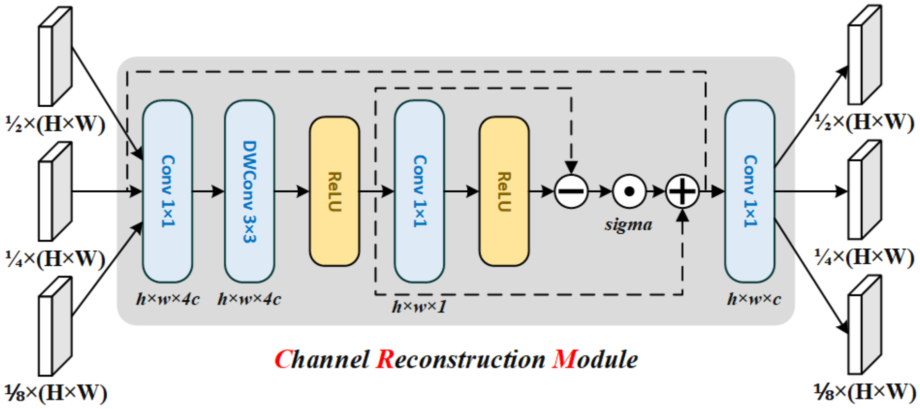

Figure 3.
Structure of the Channel Reconstruction Module.

In the hybrid U-Net network constructed in this study, the convolutional feature pre-extraction stage possesses the ability for two-dimensional local perception and fine-grained spatial structure information capture. However, it also suppresses the interaction of channel information and the extraction of key information at low resolution. This is primarily due to the continuous convolution and pooling operations during the encoding stage, which lead to the loss of spatial and boundary information. Moreover, traditional convolutions are ineffective in isolating crucial feature channels, resulting in the loss of robustness and stability in complex and noisy tasks such as cell cluster segmentation.

To address these issues, we have designed a feature channel reconstruction module (CRM) at the backend of the encoding stage at different scales. As shown in Figure 3, the CRM consists of two main stages. In the first stage, we have developed a lightweight channel expansion module to adaptively expand the high-dimensional hidden space along the channel dimension, increasing the original input channel count c to $4c$. This stage can be represented as:

$$S_1(x) = \delta_{Re\,LU}\left(\varphi_{3\times3}\left(Conv_{1\times1}^{dim=4c}(x)\right)\right) \quad (1)$$

Here, $\varphi_{3\times3}$ represents the use of depthwise separable convolutions with a kernel size of 3, which effectively reduces the computational parameters in a multi-channel design.

In the second stage, the channels will be aggregated and reconstructed. Since the channel expansion in the first stage introduces significant redundant information, although the additional hidden channels provide more nonlinear representations, they lead to an exponential increase in computational cost in subsequent stages. Therefore, we perform channel reduction and projection in this stage, reallocating the features. First, the input features are mapped to a one-dimensional plane through convolution, compressing all channels to build an efficient global representation. Then, we compute the difference between the input features and the global representation to create a more effective representation for channel scaling. Finally, we construct a residual structure and use an MLP to restore the channel dimension to c:

$$S_2(x) = Conv_{1\times1}^{dim=c}\left(x + \gamma_c \odot (x - \delta_{Re\,LU}(Conv_{1\times1}^{dim=1}(x))\right) + x \tag{2}$$

Here, γ_c represents the learnable scaling factor, which redistributes the channels of the differential features through complementary interactions during the training process.

Finally, we connect the two stages, and the constructed CRM module is a composite function of the two stages:

$$CRM(x) = S_2(S_1(x)) \tag{3}$$

The CRM module maintains dimensional consistency before and after the input, reconstructing local texture information and complex global features at each convolutional stage. It eliminates redundant channels at a relatively low additional cost.

3.4. Self-Attention Calibration Bridging Module

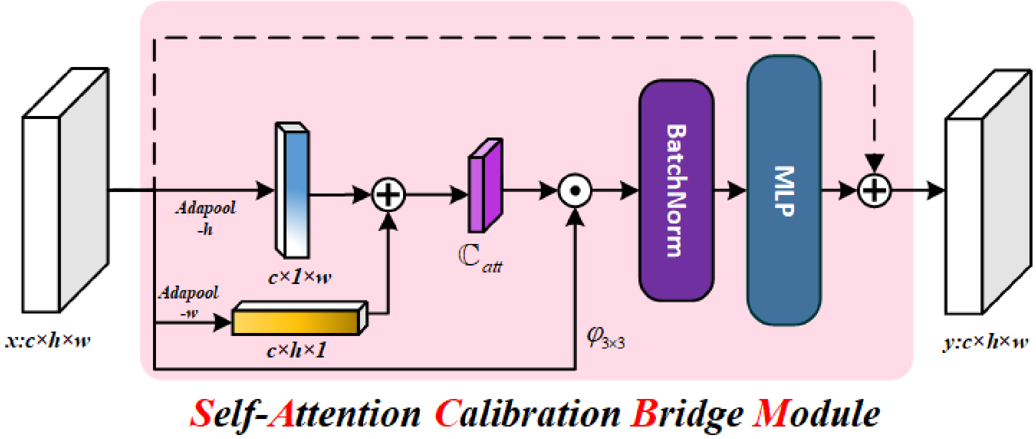

Figure 4.
Structure of the Self-Attention Calibration Bridging Module.

The U-Net framework is centered on directly transferring low-level features from the encoder to the decoder, which effectively preserves spatial information. However, this also introduces noise and irrelevant information from the cell cluster images. Skip connections do not account for the differences between low-level and high-level information. Direct concatenation and summation of low-level information with high-level features in the decoder cause the model to fail to fully leverage the differences across different layers and dimensions, particularly in complex scene information fusion processes, which leads to the loss of fine-grained semantic information and redundant noise.

To address this, we introduce a self-attention calibration bridging module (SACBM) in the skip connection process between the encoder and decoder. SACBM enhances the focus on texture features and foreground regions by capturing long-range dependencies across different axes to establish global contextual attention. The detailed structure of SACBM is shown in Figure 4. First, for the input feature xxx, adaptive mean pooling is applied in both horizontal and vertical directions to obtain two axial vectors. These vectors are then summed to generate the initial global feature with coarse attention. Subsequently, two bar convolutions, operating in different directions, are applied to decouple and calibrate the attended regions along the respective axes. The bar convolutions adaptively learn the appropriate weights in the horizontal and vertical directions to optimize the coarse attention features, forming finer attention that directs the focus toward the regions of interest for segmentation. The attention map generation process of SACBM is as follows:

$$_{att}(x) = \delta(\varphi_{k\times1}(\theta_{BN}(\varphi_{1\times k}(P_h(x) + P_w(x))))) \qquad (4)$$

Here, $\varphi_{1\times k}$ and $\varphi_{k\times1}$ represent the bar convolutions with different shapes, constructed using a large kernel depthwise separable convolution. Depthwise separable convolution decomposes the convolution operation into channel-wise and pointwise convolutions, thereby optimizing the convolution computation requirements. P_h and P_w denote the adaptive mean pooling operations in the horizontal and vertical directions, respectively. θ_{BN} represents the batch normalization operation, and δ is the sigmoid activation function.

Subsequently, the fine attention features are fused with the input features. We further extract the texture details of the input features through a 3x3 depthwise separable convolution. The fine attention features are weighted and fused with the extracted features through a dot product operation, capturing fine-grained local textures while reducing the computational parameter count. This process can be expressed as:

$$\tau(x) = \varphi_{3\times3}(x) \odot \mathbb{C}_{att}(x) \qquad (5)$$

Here, $\varphi_{3\times3}$ represents the 3x3 kernel depthwise separable convolution, and $\odot$ denotes the Hadamard product operation.

Finally, the calibrated features are further refined through batch normalization and an MLP, while residual connections are used to reuse the input features, preventing gradient explosion and vanishing gradients. The overall structure of SACBM can be expressed as follows:

$$SACBM(x) = MLP(\theta_{BN}(\tau(x))) + x \qquad (6)$$

4. Experiments

4.1. Dataset

The SIPaKMeD dataset consists of 966 cell cluster images captured with a CCD camera under an optical microscope, with certain cells manually annotated, totaling 4,049 independent labeled positive samples. These cell cluster images can be classified into five categories based on morphological and pathological features. Normal cells are divided into superficial-intermediate and parabasal categories, while abnormal cells include koilocytes and dyskeratotic cells, and benign cells are classified as metaplastic. The labeled samples categorize the cells' semantics into the nucleus, cytoplasm, and background. This study conducts experiments on these three categories for validation.

4.2. Implementation Details

The CNN used in this study is based on the ResNet-50 architecture, with weights pretrained on ImageNet and followed by 12 transformer layers at the output. For the input cell images, we apply additional data augmentation through random rotations, scaling, zooming, and flipping, and set the input image size to 512x512 pixels. All experiments were conducted on a single Nvidia RTX 3090, and during training, the SGD optimizer was used with an initial learning rate of 0.01, weight decay of 1×10−4, and a momentum factor of 0.9. For the small-scale cell dataset, the batch size was set to 4, with a total of 200 training epochs. The optimal training weights were saved using early stopping during the training process.

4.3. Quantitative Results

In this section, we present the quantitative comparison results between the proposed method and other representative baseline methods. We selected several CNN-based segmentation models, such as U-Net [40] PSPNet [41] and DDRNet [42] as well as transformer-based models like Segmenter [43]

DOI: 10.55214/25768484.v9i5.7679

and the hybrid architecture TransUNet [44]. We evaluated the performance using various metrics, including per-class IoU, mIoU, per-class Dice, mDice, per-class recall, and pixel accuracy. Among these metrics, IoU and Dice are the primary evaluation criteria, as they effectively reflect False Positive and True Negative situations in segmentation results.

As shown in Table 1, Segmenter achieved the lowest performance compared to other methods, particularly for the nuclear class. This is mainly due to the reliance of pure transformer-based models on the linear sequence of multi-head attention mechanisms, which emphasize global features while neglecting local features that are crucial for accurate nuclear segmentation. In contrast, CNN-based segmentation models performed well in terms of per-class IoU, with PSPNet achieving the best metrics for cytoplasm and background segmentation. However, its nuclear segmentation performance lagged behind TransUNet. This is because PSPNet's pyramid pooling method captures global semantic information, but the lack of skip connections in the encoding stage leads to neglect of the localization information necessary for nuclear segmentation. The proposed method outperforms others in nuclear segmentation, surpassing TransUNet by 2.51%. Although it slightly lags behind PSPNet in other categories, it achieves superior mIoU, reflecting the balanced performance of our method across different categories. The Dice metrics in Table 2 are consistent with the IoU results, further validating the segmentation performance.

In Table 3, the recall and pixel accuracy (PA) metrics show that our method underperforms compared to PSPNet. The primary reason for this is the recall calculation method, which does not account for True Negative cases. This is particularly evident in the nuclear class, where erroneous segmentation and irregular results lead to a larger proportion of True Negatives. This causes a significant discrepancy between recall and IoU. However, the area of True Negatives in the nuclear class is small and does not significantly affect the overall PA or other categories, although it still represents an overlooked error. Furthermore, PA does not consider the balance across all categories. Background and some cytoplasm classes, with larger area proportions, naturally yield higher PA scores, which disadvantages smaller area classes like nuclei. Therefore, PA is less reliable than IoU and Dice in reflecting segmentation performance for small area classes.

Table 1.
Experimental results of the proposed method and other segmentation algorithms on the IoU metric for each category (%)

Methods	other	cyt	nuc	mIoU
u-net	80.94	55.52	31.92	56.13
segmenter	80.23	51.38	8.85	46.82
ddrnet	78.46	54.07	25.49	52.67
pspnet	82.86	58.52	36.68	59.36
transunet	80.25	55.32	37.93	57.83
ours	81.68	57.96	40.44	60.03

Table 2.
Experimental results of the proposed method and other segmentation algorithms on the Dice metric for each category (%)

Methods	other	cyt	nuc	mdice
u-net	89.47	71.40	48.39	71.90
segmenter	89.03	67.88	16.26	63.78
ddrnet	87.93	70.19	40.62	69.00
pspnet	90.63	73.83	53.67	74.50
transunet	89.04	71.23	55.00	73.28
ours	89.92	73.39	57.59	75.02

DOI: 10.55214/25768484.v9i5.7679

Table 3.
Experimental results of the proposed method and other segmentation algorithms on the Recall and PA metrics for each category (%)

Methods	other	cyt	nuc	PA
u-net	82.91	79.51	57.81	83.36
segmenter	81.67	76.67	58.81	82.19
ddrnet	86.04	75.26	47.23	81.26
pspnet	83.87	87.25	67.38	85.24
transunet	83.02	78.91	64.83	83.07
ours	83.86	80.66	65.50	84.34

4.4. Qualitative Results

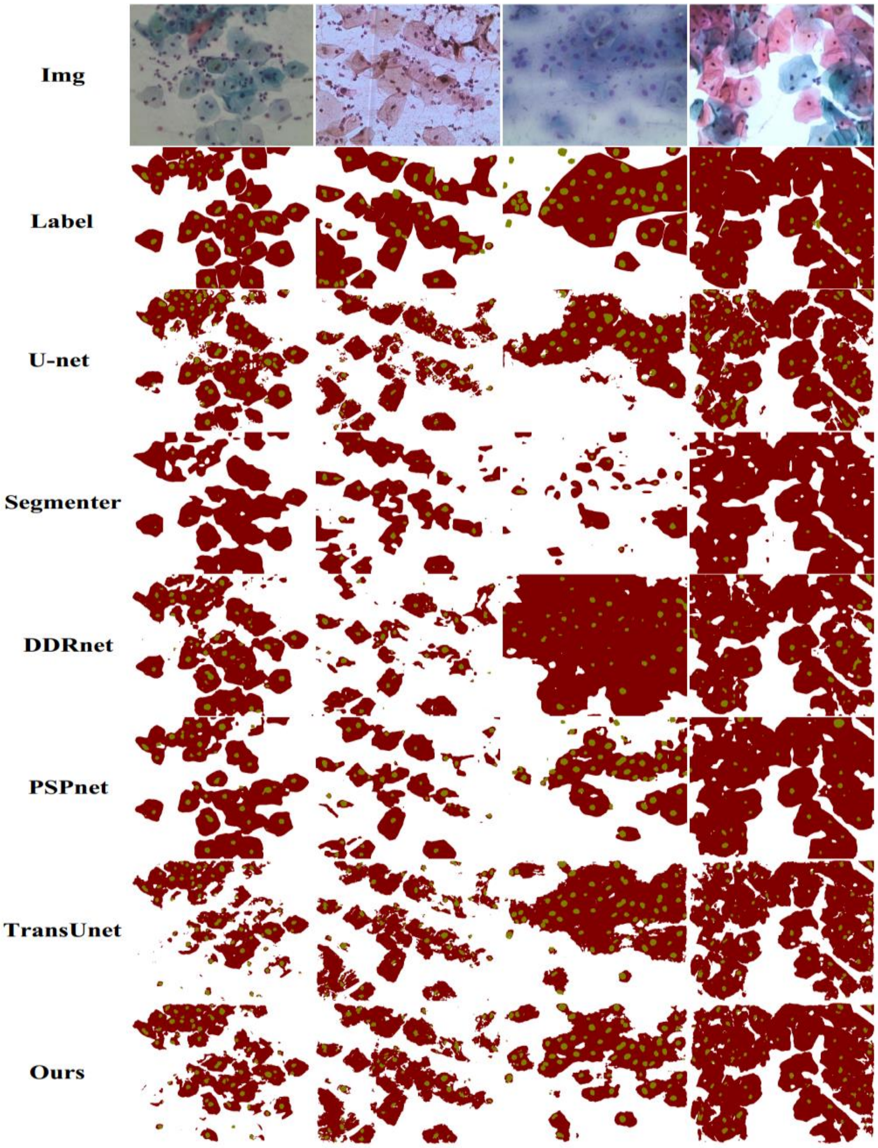

Figure 5.
The visual effectiveness of the proposed method is compared with other methods through visualization.

Figure 5 illustrates the visual performance of the proposed method compared to five other baseline methods, validating the quantitative results presented in Section 4.3. Segmenter is able to segment some general regions of the cytoplasm but fails to accurately identify the nucleus, losing significant detail due to its reliance on global inference. The U-Net model achieves good segmentation results but struggles with excessive interference, often mistakenly identifying non-cytoplasm impurities as nuclei. DDRNet, in the third sample group, segments most of the background areas as cellular regions, indicating that the multi-resolution backbone is insensitive to the blurred boundaries of the main regions. PSPNet shows better visualization across various cell types, with the cytoplasm and nucleus having correct shapes and smooth boundaries. However, its segmentation area is more conservative, suggesting that high-dimensional information is lost at the segmentation endpoint, resulting in errors due to the loss of essential cellular texture details. TransUNet retains basic texture information and maintains the main segmentation areas well, but it introduces unnecessary adhesion in the cytoplasm and fails to accurately delineate the boundaries between different cells. The proposed method outperforms TransUNet, particularly in the third and fourth samples, achieving segmentation results closer to the ground truth labels. While some impurity interference remains outside the cells, the proposed method demonstrates better robustness and fitting ability compared to U-Net and TransUNet.

4.5. Ablation Study

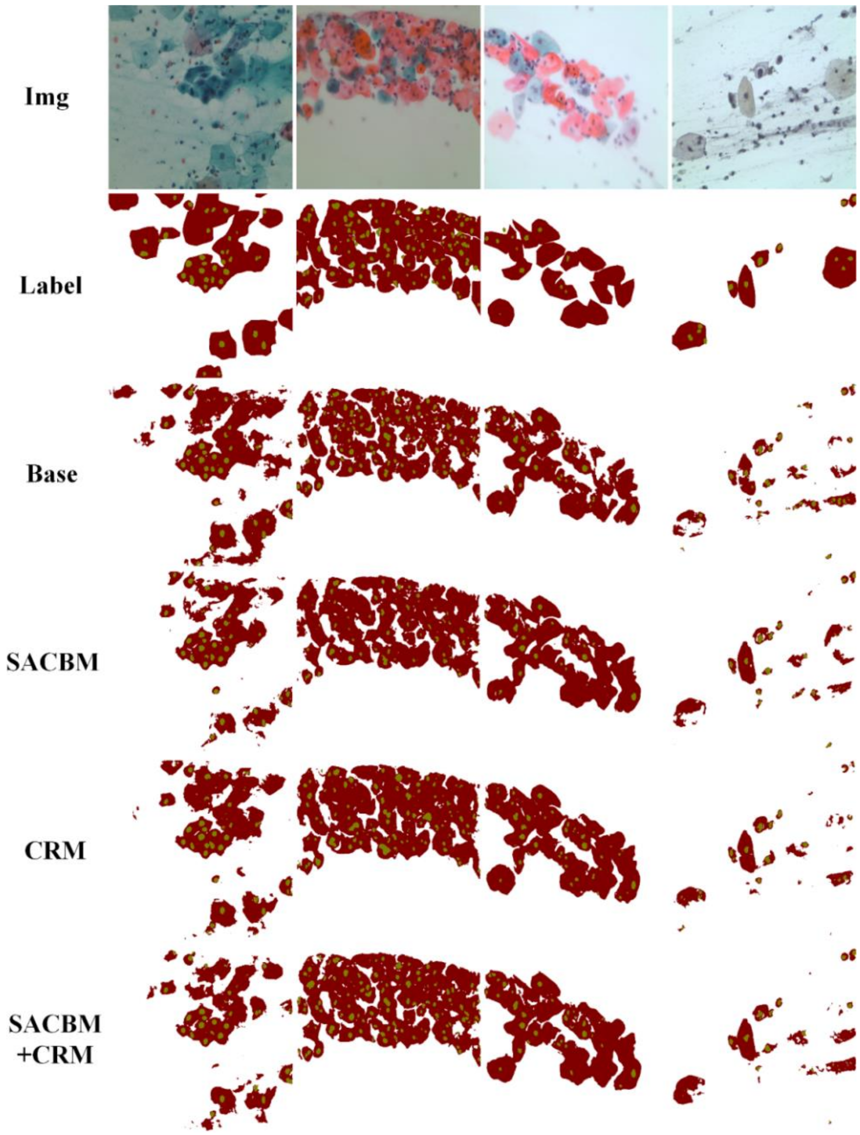

Figure 6.
Qualitative comparison of the effectiveness of SACBM and CRM through visualization methods.

Table 4.
Experimental results of the CRM and SACBM modules on the IoU metric for each category (%)

CRM	SACBM	other	cyt	nuc	mIoU
×	×	80.25	55.32	37.93	57.83
×	√	81.09	55.53	39.72	58.78
√	×	81.21	56.15	39.45	58.93
√	√	81.68	57.96	40.44	60.03

Table 5.
Experimental results of the CRM and SACBM modules on the Dice metric for each category (%)

CRM	SACBM	other	cyt	nuc	mDice
×	×	89.04	71.23	55.00	73.28
×	√	89.56	71.41	56.86	74.04
√	×	89.63	71.92	56.58	74.16
√	√	89.92	73.39	57.59	75.02

Table 6.
Experimental results of the CRM and SACBM modules on the Recall and Accuracy metrics for each category (%)

CRM	SACBM	other	cyt	nuc	PA
×	×	83.02	78.91	64.83	83.07
×	√	82.71	80.70	67.94	83.64
√	×	83.33	80.78	64.57	83.76
√	√	83.86	80.66	65.50	84.34

We conducted ablation experiments on the two core components of the proposed method, CRM and SACBM, to validate their effectiveness. Table 4 presents the ablation results of both modules on the IoU metric for each category. It can be observed that SACBM effectively improves the segmentation performance of the nucleus, primarily because the recognition of the nucleus relies on the skip connections, which preserve low-dimensional information. SACBM eliminates redundant low-dimensional information, such as impurities similar to the nucleus, during the recalibration of attention. Additionally, it retains the nucleus semantic information in the cytoplasm by leveraging global context. Similar performance improvements are observed in the Dice metric in Table 5 and the Recall metric in Table 6. On the other hand, the CRM module optimizes the performance of all features by channel compression reconstruction in the first three CNN stages, leading to consistent improvements in the IoU for all three categories. The final mIoU metric shows a 1.1% improvement over the baseline model. The combined result of both modules inherits the advantages of each, outperforming the individual use of either module across all metrics and achieving the best mIoU score of 60.03%.

The differences observed in some categories between the Recall and IoU metrics in Table 6 are primarily due to the fact that Recall does not consider false negatives (FN), i.e., missed detections. Therefore, IoU and Dice metrics are more representative compared to Recall. Nevertheless, the final combined Recall metric still outperforms the baseline model, demonstrating the effectiveness of the proposed method.

The visualizations in Figure 6 also confirm the quantitative results. The baseline model retains more impurities, aside from the cytoplasm and nucleus, leading to more fragmented and damaged segmentation in the cell clusters. The addition of SACBM improves the continuity of cell clusters, providing stronger contextual representation, while the segmentation of the nucleus becomes more consistent and complete. In the case of CRM being added individually, noise and misidentification interference are noticeably reduced compared to the baseline model, with improvements in the completeness and semantic accuracy of the cytoplasm. The combined results of SACBM and CRM achieve the closest segmentation to the ground truth, maintaining correct localization relationships and segmentation areas for the main body of the cell clusters.

5. Conclusion

This paper addresses the irregularity and excessive impurity interference in cell cluster images, proposing a Hybrid U-net architecture based on encoding channel reconstruction and self-attention calibration bridging to achieve weakly supervised extraction of cell category information in incomplete labeled cell segmentation tasks. We first performed data reconstruction on incomplete labels, transforming the weakly supervised problem into a strongly supervised one, which enhanced the impact of positive samples from prior knowledge on the segmentation performance of the network model. Additionally, we developed a channel reconstruction module in the encoding phase and a self-attention calibration bridge module to eliminate feature redundancy in the convolutional stage and to pre-attend to foreground objects in shallow features, significantly improving the semantic understanding of low-dimensional information. The proposed method was tested and validated on the publicly available weakly supervised cell segmentation dataset SIPaKMeD. Through comparative experiments and ablation studies, the modules proposed in this paper achieved excellent performance in both quantitative metrics and visual analysis, outperforming other semantic segmentation networks in terms of mIoU, Dice, and other indicators.

Funding:

This research was funded by Anhui Province Quality Engineering Project grant number2020mooc194 and Anhui Province Quality Engineering Project grant number 2024cxtdx02

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Acknowledgments:

I also wish to thank my family for their unwavering support and encouragement. Their love and understanding have been a constant source of strength throughout this journey.

Copyright:

References

[1] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022. https://doi.org/10.1109/LGRS.2022.3144214

[2] A. Froidure *et al.*, "Correlation of BAL cell count and pulmonary function tests in the era of antifibrotics: data from the Belgium-Luxembourg idiopathic pulmonary fibrosis registry," *Chest*, vol. 163, no. 2, pp. 358-361, 2023. https://doi.org/10.1016/j.chest.2022.11.007

[3] S. Shafi and A. V. Parwani, "Artificial intelligence in diagnostic pathology," *Diagnostic pathology*, vol. 18, no. 1, p. 109, 2023. https://doi.org/10.1186/s13000-023-01356-3

[4] J. Feng, J. Neuzil, A. Manz, C. Iliescu, and P. Neuzil, "Microfluidic trends in drug screening and drug delivery," *TrAC Trends in Analytical Chemistry*, vol. 158, p. 116821, 2023. https://doi.org/10.1016/j.trac.2023.116821

[5] P. L. Kavanagh, T. A. Fasipe, and T. Wun, "Sickle cell disease: A review," *Jama*, vol. 328, no. 1, pp. 57-68, 2022. https://doi.org/10.1001/jama.2022.10224

[6] L. Abualigah, K. H. Almotairi, and M. A. Elaziz, "Multilevel thresholding image segmentation using meta-heuristic optimization algorithms: Comparative analysis, open challenges and new trends," *Applied Intelligence*, vol. 53, no. 10, pp. 11654-11704, 2023. https://doi.org/10.1007/s10462-023-02216-9

[7] J. Jing, S. Liu, G. Wang, W. Zhang, and C. Sun, "Recent advances on image edge detection: A comprehensive review," *Neurocomputing*, vol. 503, pp. 259-271, 2022. https://doi.org/10.1016/j.neucom.2022.08.065

[8] R. Sun *et al.*, "Survey of image edge detection," *Frontiers in Signal Processing*, vol. 2, p. 826967, 2022. https://doi.org/10.3389/fsigp.2022.826967

[9] N. S. M. Raja, S. L. Fernandes, N. Dey, S. C. Satapathy, and V. Rajinikanth, "Contrast enhanced medical MRI evaluation using Tsallis entropy and region growing segmentation," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-12, 2024. https://doi.org/10.1007/s12652-024-05111-2

[10] F. Oktavia, D. Andreswari, and B. Susilo, "Segmentasi Citra Diaretdb1 Pada Area hemorrhages diabetic retinopathy Menggunakan Metode Region growing," *Rekursif: Jurnal Informatika*, vol. 10, no. 1, pp. 1-11, 2022.

[11] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, 2023. https://doi.org/10.1109/TGRS.2023.3235682

[12] P. Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. ul Haq, "Understanding of convolutional neural network (cnn): A review," *International Journal of Robotics and Control Systems*, vol. 2, no. 4, pp. 739-748, 2022. https://doi.org/10.1007/s43235-022-00033-3

[13] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364-376, 2022. https://doi.org/10.1109/TIP.2022.3150289

[14] S. Castro, V. Pereira, and R. Silva, "Improved segmentation of cellular Nuclei using UNET architectures for enhanced pathology imaging," *Electronics*, vol. 13, no. 16, p. 3335, 2024. https://doi.org/10.3390/electronics13163335

[15] X. Xu *et al.*, "Crack detection and comparison study based on faster R-CNN and mask R-CNN," *Sensors*, vol. 22, no. 3, p. 1215, 2022. https://doi.org/10.3390/s22041215

[16] X. Bi, J. Hu, B. Xiao, W. Li, and X. Gao, "Iemask r-cnn: Information-enhanced mask r-cnn," *IEEE Transactions on Big Data*, vol. 9, no. 2, pp. 688-700, 2022. https://doi.org/10.1109/TBDATA.2022.3173149

[17] K. Löffler and R. Mikut, "EmbedTrack—Simultaneous cell segmentation and tracking through learning offsets and clustering bandwidths," *IEEE Access*, vol. 10, pp. 77147-77157, 2022. https://doi.org/10.1109/ACCESS.2022.3183951

[18] Q. Wang, J. Wei, and B. Quan, "Regionally adaptive active learning framework for nuclear segmentation in microscopy image," *Electronics*, vol. 13, no. 17, p. 3430, 2024. https://doi.org/10.3390/electronics13173430

[19] X. Huo *et al.*, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, p. 105534, 2024. https://doi.org/10.1016/j.bspc.2023.105534

[20] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331-368, 2022. https://doi.org/10.1007/s41095-022-0215-0

[21] W. Shen *et al.*, "A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9284-9305, 2023. https://doi.org/10.1109/TPAMI.2023.3296019

[22] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-based semi-supervised learning: A comprehensive review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8174-8194, 2022. https://doi.org/10.1109/TNNLS.2022.3205156

[23] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *Journal of Big Data*, vol. 9, no. 1, p. 102, 2022. https://doi.org/10.1186/s40537-022-00629-2

[24] M. Veta, P. J. Van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in H&E stained breast cancer histopathology images," *PloS one*, vol. 8, no. 7, p. e70221, 2013. https://doi.org/10.1371/journal.pone.0070221

[25] S. Wienert *et al.*, "Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach," *Scientific Reports*, vol. 2, no. 1, p. 503, 2012. https://doi.org/10.1038/srep00503

[26] M. Liao *et al.*, "Automatic segmentation for cell images based on bottleneck detection and ellipse fitting," *Neurocomputing*, vol. 173, pp. 615-622, 2016. https://doi.org/10.1016/j.neucom.2015.07.087

[27] C. Li *et al.*, "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," presented at the 2020 IEEE International Conference on Image Processing (ICIP), 2020.

[28] K. Fan, S. Wen, and Z. Deng, "Deep learning for detecting breast cancer metastases on WSI," in *Innovation in Medicine and Healthcare Systems, and Multimedia: Proceedings of KES-InMed-19 and KES-IIMSS-19 Conferences*, 2019: Springer, pp. 137-145.

[29] T. Tran, O.-H. Kwon, K.-R. Kwon, S.-H. Lee, and K.-W. Kang, "Blood cell images segmentation using deep learning semantic segmentation," presented at the 2018 IEEE International Conference on Electronics and Communication Engineering (ICECE), 2018.

[30] A. Mahbod *et al.*, "Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images," *Scientific Data*, vol. 11, no. 1, p. 295, 2024. https://doi.org/10.1038/s41597-024-02349-1

[31] L. Wang, D. Guo, G. Wang, and S. Zhang, "Annotation-efficient learning for medical image segmentation based on noisy pseudo labels and adversarial learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2795-2807, 2020.

[32] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, "Affinity attention graph neural network for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8082-8096, 2021.

[33] B. Chen, A. Rouditchenko, K. Duarte, and H. Kuehne, "Multimodal clustering networks for self-supervised learning from unlabeled videos," in *Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8012–8021.

[34] H. Du, H. Shi, P. Zhao, and D. Wang, "Contrastive learning with bidirectional transformers for sequential recommendation," in *Proceedings of the Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 396–405.

[35] W. Xu, H. Huang, and M. Cheng, "Masked cross-image encoding for few-hot segmentation," in *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME) IEEE*, 2023, pp. 744–749.

[36] J. Zhu *et al.*, "Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation," *arXiv preprint arXiv:2312.09251*, 2023.

DOI: 10.55214/25768484.v9i5.7679

[37] Y. Lin, M. Chen, W. Wang, and B. Wu, "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," in *Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15305–15314.

[38] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1-10, 2011.

[39] I. S. Kwon *et al.*, "Two-dimensional MoS 2/Fe-phthalocyanine hybrid nanostructures as excellent electrocatalysts for hydrogen evolution and oxygen reduction reactions," *Nanoscale*, vol. 11, no. 30, pp. 14266-14275, 2019.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention– MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer*, 2015, pp. 234–241.

[41] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[42] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *arXiv preprint arXiv:2101.06085*, 2021.

[43] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[44] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.