TransNuSeg: A Lightweight Multi-task Transformer for Nuclei Segmentation

Zhenqi He[1], Mathias Unberath[2], Jing Ke[3], and Yiqing Shen[2(✉)]

[1] The University of Hong Kong, Pok Fu Lam, Hong Kong
[2] Johns Hopkins University, Baltimore, USA
yshen92@jhu.edu
[3] Shanghai Jiao Tong University, Shanghai, China

Abstract. Nuclei appear small in size, yet, in real clinical practice, the global spatial information and correlation of the color or brightness contrast between nuclei and background, have been considered a crucial component for accurate nuclei segmentation. However, the field of automatic nuclei segmentation is dominated by Convolutional Neural Networks (CNNs), meanwhile, the potential of the recently prevalent Transformers has not been fully explored, which is powerful in capturing local-global correlations. To this end, we make the first attempt at a pure Transformer framework for nuclei segmentation, called `TransNuSeg`. Different from prior work, we decouple the challenging nuclei segmentation task into an intrinsic multi-task learning task, where a tri-decoder structure is employed for nuclei instance, nuclei edge, and clustered edge segmentation respectively. To eliminate the divergent predictions from different branches in previous work, a novel self distillation loss is introduced to explicitly impose consistency regulation between branches. Moreover, to formulate the high correlation between branches and also reduce the number of parameters, an efficient attention sharing scheme is proposed by partially sharing the self-attention heads amongst the tri-decoders. Finally, a token MLP bottleneck replaces the over-parameterized Transformer bottleneck for a further reduction in model complexity. Experiments on two datasets of different modalities, including MoNuSeg have shown that our methods can outperform state-of-the-art counterparts such as CA$^{2.5}$-Net by 2–3% Dice with 30% fewer parameters. In conclusion, `TransNuSeg` confirms the strength of Transformer in the context of nuclei segmentation, which thus can serve as an efficient solution for real clinical practice. Code is available at https://github.com/zhenqi-he/transnuseg.

Keywords: Lightweight Multi-Task Framework · Shared Attention Heads · Nuclei · Edge and Clustered Edge Segmentation

1 Introduction

Accurate cancer diagnosis, grading, and treatment decisions from medical images heavily rely on the analysis of underlying complex nuclei structures [7]. Yet, due

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 206–215, 2023.
https://doi.org/10.1007/978-3-031-43901-8_20

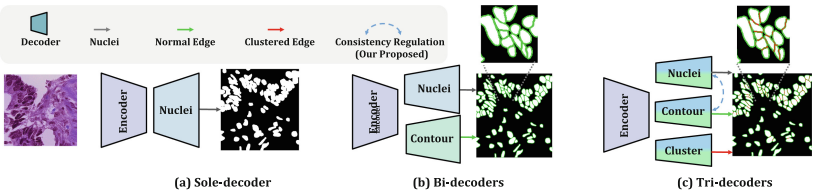

Fig. 1. Semantic illustrations of the nuclei segmentation networks with different numbers of decoders. (a) Sole-decoder to perform a single task of nuclei segmentation. (b) Bi-decoder to segment nuclei and locate nuclei edges simultaneously. (c) Tri-decoder with the third encoder path to specify the challenging clustered edge (ours), where the consistency regularization is designed across the predictions from the other two branches (dashed line).

to the numerous nuclei contained in a digitized whole-slide image (WSI), or even in an image patch of deep learning input, dense annotation of nuclei contouring is extremely time-consuming and labor-expensive [11]. Consequently, automated nuclei segmentation approaches have emerged to satisfy a broad range of computer-aided diagnostic systems, where the deep learning methods, particularly the convolutional neural networks [5,12,14,19,21] have received notable attention due to their simplicity and generalization ability.

In the literature work, the sole-decoder design in these UNet variants (Fig. 1(a)) is susceptible to failures in splitting densely clustered nuclei when precise edge information is absent. Hence, deep contour-aware neural network (DCAN) [3] with bi-decoder structure achieves improved instance segmentation performance by adopting multi-task learning, in which one decoder learns to segment the nuclei and the other recognizes edges as described in Fig. 1(b). Similarly, CIA-Net [20] extends DCAN with an extra information aggregator to fuse the features from two decoders for more precise segmentation. Much recently, CA$^{2.5}$-Net [6] shows identifying the clustered edges in a multiple-task learning manner can achieve higher performance, and thereby proposes an extra output path to learn the segmentation of clustered edges explicitly. A significant drawback of the aforementioned multi-decoder networks is the ignorance of the prediction consistency between branches, resulting in sub-optimal performance and missing correlations between the learned branches. Specifically, a prediction mismatch between the nuclei and edge branches is observed in previous work [8], implying a direction for performance improvement. To narrow this gap, we propose a consistency distillation between the branches, as shown by the dashed line in Fig. 1(c). Furthermore, to resolve the cost of involving more decoders, we propose an attention sharing scheme, along with an efficient token MLP bottleneck [16], which can both reduce the number of parameters.

Additionally, existing methods are CNN-based, and their intrinsic convolution operation fails to capture global spatial information or the correlation amongst nuclei [18], which domain experts rely heavily on for accurate nuclei allocation. It suggests the presence of long-range correlation in practical nuclei

segmentation tasks. Inspired by the capability in long-range global context capturing by Transformers [17], we make the first attempt to construct a tri-decoder based Transformer model to segment nuclei. In short, our major contributions are three-fold: (1) We propose a novel multi-task framework for nuclei segmentation, namely `TransNuSeg`, as the first attempt at a fully Swin-Transformer driven architecture for nuclei segmentation. (2) To alleviate the prediction inconsistency between branches, we propose a novel self distillation loss that regulates the consistency between the nuclei decoder and normal edge decoder. (3) We propose an innovative attention sharing scheme that shares attention heads amongst all decoders. By leveraging the high correlation between tasks, it can communicate the learned features efficiently across decoders and sharply reduce the number of parameters. Furthermore, the incorporation of a light-weighted MLP bottleneck leads to a sharp reduction of parameters at no cost of performance decline.

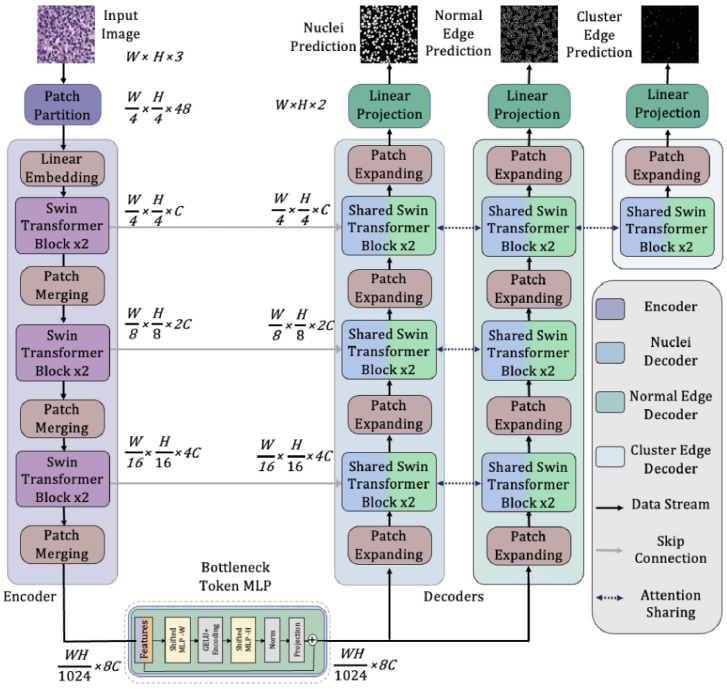

Fig. 2. The overall framework of the proposed `TransNuSeg` of three output branches to separate the nuclei, normal edges, and cluster edges, respectively. In the novel design, a pre-defined proportion of the attention heads are shared between the decoders via the proposed sharing scheme, which considerably reduces the number of parameters and enables more efficient information communication.

2 Methodology

Network Architecture Overview. Figure 2 illustrates the overall architecture of the proposed multi-task tri-decoder Transformer network, named `TransNuSeg`. Both the encoder and decoders utilize the Swin Transformer [13] as the building blocks to capture the long-range feature correlations in the nuclei segmentation context. Our network consists of three individual output decoder paths for nuclei segmentation, normal edges segmentation, and clustered edges segmentation. Given the high dependency between edge and clustered edge, we are inspired to propose a novel attention sharing scheme, which can communicate the information and share learned features across decoders while also reducing the number of parameters. Additionally, a token MLP bottleneck is incorporated to further increase the model efficiency.

Attention Sharing Scheme. To capture the strong correlation between nuclei segmentation and contour segmentation between multiple decoders [15], we introduce a novel attention sharing scheme that is designed as an enhancement to the multi-headed self-attention (MSA) module in the plain Transformer [17]. Based on the attention sharing scheme, we design a shared MSA module, which is similar in structure to vanilla MSA. Specifically, it consists of a `LayerNorm` layer [1], residual connection, and feed-forward layer. Innovatively, it differs from the vanilla MSA by sharing a proportion of globally-shared self-attention (SA) heads amongst all the parallel Transformer blocks in decoders, while keeping the remaining SA heads unshared *i. e.* learn the weights separately. A schematic illustration of the shared MSA module in the Swin Transformer block is demonstrated in Fig. 3, as is formally formulated as follows:

$$\text{Shared-MSA}(\mathbf{z}) = \left[\text{SA}_1^s(\mathbf{z}), \cdots, \text{SA}_m^s(\mathbf{z}), \text{SA}_1^u(\mathbf{z}), \cdots, \text{SA}_n^u(\mathbf{z})\right] \mathbf{U}_{\text{MSA}}^u, \tag{1}$$

$[\cdot]$ writes for the concatenation, $\text{SA}(\cdot)$ denotes the self-attention head whose output dimension is D_h, and $\mathbf{U}_{\text{MSA}}^u \in \mathbb{R}^{(m+n)\cdot D_h \times D}$ is a learnable matrix. The superscript s and u refer to the globally-shared and unshared weights across all decoders, respectively.

Token MLP Bottleneck. To reduce the complexity of the model, we leverage a token MLP bottleneck as a light-weight alternative for the Swin Transformer bottleneck. Specifically, this approach involves shifting the latent features extracted by the encoder via two MLP blocks across the width and height channels, respectively [16]. The objective of this process is to attend to specific areas, which mimics the shifted window attention mechanism in Swin Transformer [13]. The shifted features are then projected by a learnable MLP and normalized through a `LayerNorm` [1] before being fed to a reprojection MLP layer.

Consistency Self Distillation. To alleviate the inconsistency between the contour generated from the nuclei segmentation prediction and the predicted edge,

we propose a novel consistency self distillation loss, denoted as $\mathcal{L}_{SD}$. Formally, this regularization is defined as the dice loss between the contour generated from the nuclei branch prediction (y_n) using the Sobel operation ($\texttt{sobel}(y_n)$) and the predicted edges y_e from the normal edge decoder. Specifically, the self distillation loss $\mathcal{L}_D$ is formulated by $\mathcal{L}_{sd} = \texttt{Dice}(\texttt{sobel}(y_n), y_e)$.

Multi-task Learning Objective. We employ a multi-task learning paradigm to train the tri-decoder network, aiming to improve model performance by leveraging the additional supervision signal from edges. Particularly, the nuclei semantic segmentation is considered the primary task, while the normal edge and clustered edge semantic segmentation are viewed as auxiliary tasks. All decoder branches follow a uniform scheme that combines the cross-entropy loss and the dice loss, with the balancing coefficients set to 0.60 and 0.40 respectively, as previous work [6]. Subsequently, the overall loss $\mathcal{L}$ is calculated as a weighted summation of semantic nuclei mask loss ($\mathcal{L}_n$), normal edge loss ($\mathcal{L}_e$), and clustered edge loss ($\mathcal{L}_c$), and the self distillation loss ($\mathcal{L}_{SD}$) *i. e.*

$\mathcal{L} = \gamma_n \cdot \mathcal{L}_n + \gamma_e \cdot \mathcal{L}_e + \gamma_c \cdot \mathcal{L}_c + \gamma_{sd} \cdot \mathcal{L}_{sd}$, where coefficients γ_n, γ_e and γ_c are set to 0.30, 0.35, 0.35 respectively, and γ_{sd} is initially set to 1 with a 0.3 decrease for every 10 epochs until it reaches 0.4.

3 Experiments

Dataset. We evaluated the applicability of our approach across multiple modalities by conducting evaluations on microscopy and histology datasets. (1) *Fluorescence Microscopy Image Dataset*: This set combines three different data sources to simulate the heterogeneous nature of medical images [9]. It consists of 524 fluorescence images, each with a resolution of 512 × 512 pixels. (2) *Histology Image Dataset*: This set is the combination of the open dataset MoNuSeg [10] and another private histology dataset [8] of 462 images. We crop each image in the MoNuSeg dataset into four partially overlapping 512 × 512 images.

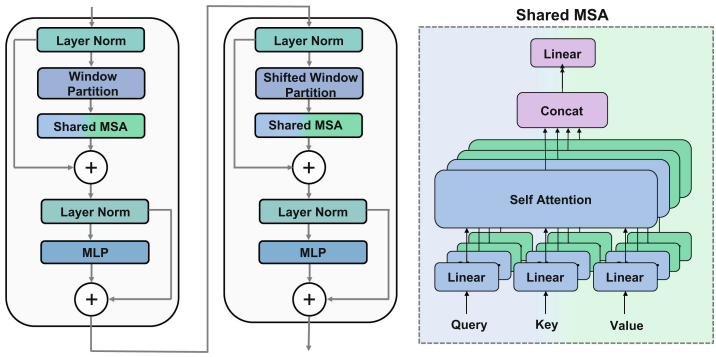

Fig. 3. A schematic illustration of the proposed Attention Sharing scheme.

Table 1. Quantitative comparisons with counterparts. The best performance with respect to each metric is highlighted in **boldface**.

Dataset	Methods	DSC (%)	F1 (%)	Acc (%)	IoU (%)	ErCnt (%)
Microscopy	UNet	85.51 ± 0.35	91.05 ± 0.13	92.19 ± 0.20	85.44 ± 0.29	55.2 ± 2.7
	UNet++	94.14 ± 0.58	92.34 ± 0.63	93.87 ± 0.61	86.20 ± 1.02	69.3 ± 1.4
	TransUNet	94.14 ± 0.47	92.31 ± 0.34	93.76 ± 0.50	86.16 ± 0.56	51.9 ± 1.0
	SwinUNet	96.05 ± 0.27	95.02 ± 0.23	96.08 ± 0.23	91.06 ± 0.43	31.2 ± 0.6
	$CA^{2.5}$-Net	91.08 ± 0.49	90.05 ± 0.27	93.40 ± 0.14	86.89 ± 0.87	18.6 ± 1.3
	Ours	**97.01** ± 0.74	**96.67** ± 0.60	**97.11** ± 1.02	**92.97** ± 0.41	**9.78** ± 2.1
Histology	UNet	80.97 ± 0.75	72.17 ± 0.49	90.14 ± 0.24	61.63 ± 0.36	45.7 ± 1.6
	UNet++	87.10 ± 0.16	75.20 ± 0.19	91.34 ± 0.14	62.89 ± 0.27	38.0 ± 2.4
	TransUNet	85.80 ± 0.20	72.87 ± 0.49	90.53 ± 0.27	60.21 ± 0.46	35.2 ± 0.8
	SwinUNet	88.73 ± 0.90	78.11 ± 1.88	91.23 ± 0.73	64.41 ± 0.15	27.6 ± 2.3
	$CA^{2.5}$-Net	86.74 ± 0.18	77.42 ± 0.30	91.52 ± 0.78	66.79 ± 0.34	23.7 ± 0.7
	Ours	**90.81** ± 0.22	**81.52** ± 0.44	**92.77** ± 0.64	**69.49** ± 0.17	**11.4** ± 1.1

The private dataset contains 300 images sized at 512 × 512 tessellated from 50 WSIs scanned at 20×, and meticulously labeled by five pathologists according to the labeling guidelines of the MoNuSeg [10]. For both datasets, we randomly split 80% of the samples on the patient level as the training set and the remaining 20% as the test set.

Table 2. Comparison of the model complexity in terms of the number of parameters, FLOPs, as well as the training cost in the form of the averaged training time per epoch. The average training time is computed using the same batch size for both datasets, with the first number indicating the averaged time on the Fluorescence Microscopy Image Dataset and the second on the Histology Image Dataset. The token MLP bottleneck and attention sharing scheme are denoted as 'MLP', and 'AS', respectively.

Methods	#Params ($\times 10^6$)	FLOPs ($\times 10^9$)	Training (s)
UNet [14]	31.04	219.03	43.4/27.7
UNet++ [21]	9.05	135.72	41.8/31.7
TransUNet [4]	67.87	129.97	37.1/34.5
SwinUNet [2]	27.18	30.67	37.8/35.2
$CA^{2.5}$-Net [6]	24.27	460.70	73.8/70.2
Ours (w/o MLP & w/o AS)	34.33	93.98	76.1/74.3
Ours (w/o MLP)	30.82	123.60	62.6/61.2
Ours (w/o AS)	21.33	116.95	53.1/51.2
Ours (full settings)	17.82	165.95	51.5/50.8

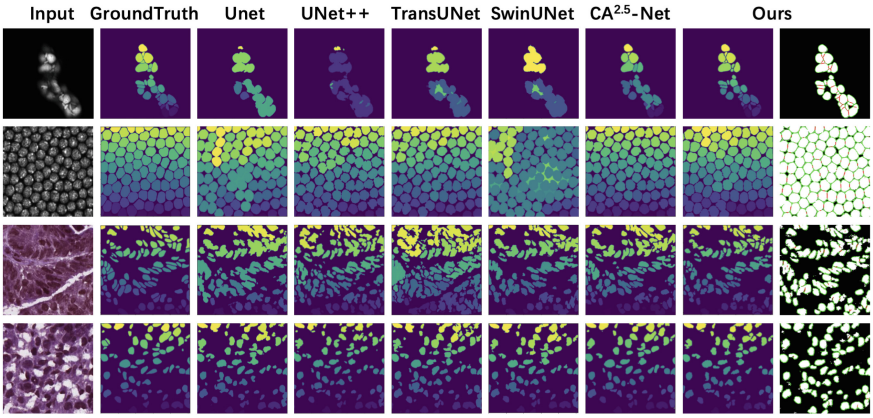

Fig. 4. Exemplary samples and their segmentation results using different methods. `TransNuSeg` demonstrates superior segmentation performance compared to its counterparts, which can successfully distinguish severely clustered nuclei from normal edges.

Implementations. All experiments are performed on one NVIDIA RTX 3090 GPU with 24 GB memory. We use Adam optimizer with an initial learning rate of 1×10^{-4}. We compare `TransNuSeg` with UNet [14], UNet++ [21], TransUNet [4], SwinUNet [2], and CA$^{2.5}$-Net [6]. We evaluate the results by using Dice Score (DSC), Intersection over Union (IoU), pixel-level accuracy (Acc), and F1-score(F1) as metrics, and ErCnt [8]. To ensure statistical significance, we run all methods five times with different fixed seeds and report the results as mean $\pm$ standard deviation.

Results. Table 1 shows the quantitative comparisons for the nuclei segmentation. The large margin between the SwinUNet and the other CNN-based or hybrid networks also confirms the superiority of the Transformer in fine-grained nuclei segmentation. More importantly, our method can outperform SwinUNet and the previous methods on both datasets. For example, in the histology image dataset, `TransNuSeg` improves the dice score, F1 score, accuracy, and IoU by 2.08%, 3.41%, 1.25%, and 2.70% respectively, over the second-best models. Similarly, in the fluorescence microscopy image dataset, our proposed model improves DSC by 0.96%, while also leading to 1.65%, 1.03% and 1.91% increment in F1 score, accuracy, and IoU to the second-best performance. For better visualization, representative samples and their segmentation results using different methods are demonstrated in Fig. 4. Furthermore, Table 2 compares the model complexity in terms of the number of parameters, floating point operations per second (FLOPs), and the training computational cost, where our approach can significantly reduce around 28% of the training time compared to the state-of-the-art CNN multi-task method CA$^{2.5}$-Net, while also boosting performance.

Table 3. The ablation on each functional block, where 'MLP', 'AS', and 'SD' represent the token MLP bottleneck, attention sharing scheme, and the self distillation.

MLP	AS	SD	Microscopy				Histology			
			DSC (%)	F1 (%)	Acc (%)	IoU (%)	DSC (%)	F1 (%)	Acc (%)	IoU (%)
×	×	×	95.31	94.05	96.06	90.05	88.76	78.20	90.96	64.48
•	×	×	95.49	94.48	95.95	89.97	89.41	77.94	91.02	65.17
×	•	×	95.88	93.51	96.11	90.55	90.23	80.46	92.03	67.84
•	•	×	96.95	95.72	96.92	91.98	90.27	81.04	92.01	67.56
×	×	•	96.99	95.74	97.02	92.22	90.25	80.81	92.45	68.14
•	×	•	96.58	95.65	97.03	92.07	90.17	80.62	92.35	67.88
×	•	•	96.89	95.78	97.12	92.08	90.34	80.88	92.49	68.05
•	•	•	**97.01**	**96.67**	**97.11**	**92.97**	**90.81**	**81.52**	**92.77**	**69.49**

Ablation. Our ablation study yields that token MLP bottleneck and attention sharing schemes can complementarily reduce the training cost while increasing efficiency, as shown in Table 2 (the last 4 rows). To further show the effectiveness of these schemes, as well as consistency self distillation, we conduct a comprehensive ablation study on both datasets. As described in Table 3, each component proportionally contributes to the improvement to reach the overall performance

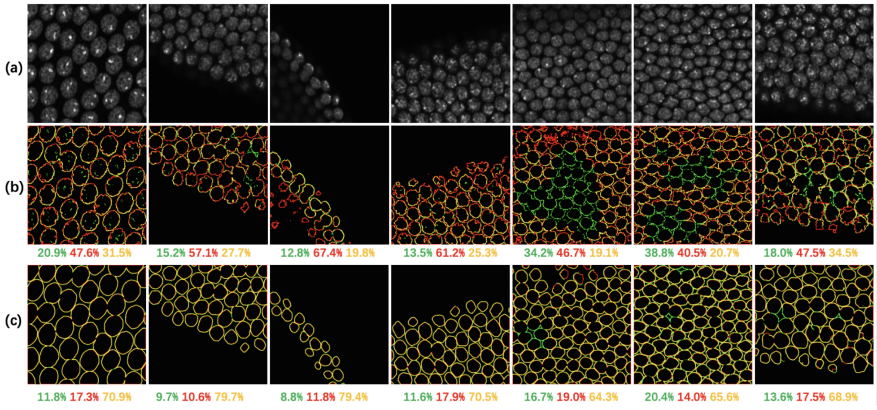

Fig. 5. The impact of self distillation regularization on mismatch reduction across three decoders. (a) Raw input image. Segmentation results by `TransNuSeg` trained (b) w/o self distillation, and (c) w/ self distillation. The predicted normal edges from the normal edge decoder are shown in green; while the edges generated from the nuclei decoder and processed with the Sobel operation are in red. The yellow color indicates the overlap between both. Accordingly, the numbers below images indicate the proportion of the pixels belonging to the three parts. Compared to the results without self distillation, the outputs with self distillation exhibit reduced mismatches, resulting in improved segmentation performance. (Color figure online)

boost. Moreover, self distillation can enhance the intrinsic consistency between two branches, as visualized in Fig. 5.

4 Conclusion

In this paper, we make the first attempt at an efficient but effective multi-task Transformer framework for modality-agnostic nuclei segmentation. Specifically, our tri-decoder framework `TransNuSeg` leverages an innovative self distillation regularization to impose consistency between the different branches. Experimental results on two datasets demonstrate the excellence of our `TransNuSeg` against state-of-the-art counterparts for potential real-world clinical deployment. Additionally, our work opens a new architecture to perform nuclei segmentation tasks with Swin Transformer, where further investigations can be performed to explore the generalizability to the top of our methods with different modalities.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cao, H., Wang, Y., Chen, J., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-25066-8_9
3. Chen, H., Qi, X., Yu, L., Heng, P.A.: DCAN: deep contour-aware networks for accurate gland segmentation. CoRR, abs/1604.02677 (2016)
4. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. CoRR, abs/2102.04306 (2021)
5. Guo, R., Pagnucco, M., Song, Y.: Learning with noise: mask-guided attention model for weakly supervised nuclei segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 461–470. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_43
6. Huang, J., Shen, Y., Shen, D., Ke, J.: $CA^{2.5}$-net nuclei segmentation framework with a microscopy cell benchmark collection. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 445–454. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_43
7. Irshad, H., Veillard, A., Roux, L., Racoceanu, D.: Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. IEEE Rev. Biomed. Eng. **7**, 97–114 (2014)
8. Ke, J., et al.: ClusterSeg: a crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. Med. Image Anal. **85**, 102758 (2023)
9. Kromp, F., et al.: An annotated fluorescence image dataset for training nuclear segmentation methods. Sci. Data **7**(1), 262 (2020)
10. Kumar, N., Verma, R., Anand, D., et al.: A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging **39**(5), 1380–1391 (2020)
11. Lagree, A., et al.: A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks. Sci. Rep. **11**(1), 8025 (2021)

12. Lal, S., Das, D., Alabhya, K., Kanfade, A., Kumar, A., Kini, J.: Nucleiseg-net: robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images. Comput. Biol. Med. **128**, 01 (2021)
13. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. CoRR, abs/2103.14030 (2021)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597 (2015)
15. Shen, Y.: Federated learning for chronic obstructive pulmonary disease classification with partial personalized attention mechanism. arXiv preprint arXiv:2210.16142 (2022)
16. Valanarasu, J.M.J., Patel, V.M.: UneXt: MLP-based rapid medical image segmentation network. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 23–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_3
17. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Von Luxburg, U., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates Inc, (2017)
18. Wang, C., Xu, R., Xu, S., Meng, W., Zhang, X.: DA-Net: dual branch transformer and adaptive strip upsampling for retinal vessels segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 528–538. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_51
19. Wazir, S., Fraz, M.M.: HistoSeg: quick attention with multi-loss function for multi-structure segmentation in digital histology images. In: 2022 12th International Conference on Pattern Recognition Systems (ICPRS). IEEE (2022)
20. Zhou, Y., Onder, O.F., Dou, Q., Tsougenis, E., Chen, H., Heng, P.A.: Cia-net: robust nuclei instance segmentation with contour-aware information aggregation. CoRR, abs/1903.05358 (2019)
21. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1