SAM Guided Task-Specific Enhanced Nuclei Segmentation in Digital Pathology

Bishal R. Swain[1], Kyung J. Cheoi[2], and Jaepil Ko[1](✉)

[1] Kumoh National Institute of Technology, Gumi, Korea 39177
{bishalswain,nonezero}@kumoh.ac.kr
[2] Chungbuk National University, Cheongju, Korea 28644
kjcheoi@chungbuk.ac.kr

Abstract. Cell nuclei segmentation is crucial in digital pathology for various diagnoses and treatments which are prominently performed using semantic segmentation that focus on scalable receptive field and multi-scale information. In such segmentation tasks, U-Net based task-specific encoders excel in capturing fine-grained information but fall short integrating high-level global context. Conversely, foundation models inherently grasp coarse-level features but are not as proficient as task-specific models to provide fine-grained details. To this end, we propose utilizing the foundation model to guide the task-specific supervised learning by dynamically combining their global and local latent representations, via our proposed X-Gated Fusion Block, which uses Gated squeeze and excitation block followed by Cross-attention to dynamically fuse latent representations. Through our experiments across datasets and visualization analysis, we demonstrate that the integration of task-specific knowledge with general insights from foundational models can drastically increase performance, even outperforming domain-specific semantic segmentation models to achieve state-of-the-art results by increasing the Dice score and mIoU by approximately 12% and 17.22% on CryoNuSeg, 15.55% and 16.77% on NuInsSeg, and 9% on both metrics for the CoNIC dataset. Our code will be released at https://cvpr-kit.github.io/SAM-Guided-Enhanced-Nuclei-Segmentation/.

Keywords: Nuclei Segmentation · Histopathology · Digital Pathology.

1 Introduction

In the domain of digital pathology, accurate segmentation of nuclei in histopathological images is critical for the research of cancer diagnosis [13]. Nuclei segmentation facilitates detailed examination of cellular behaviors, including the analysis of cell cycles and the investigation of mutations in proteins linked to cancer [9]. The precision of segmentation directly impacts the analysis of tissue biopsies, which are conducted millions of times each year and represent the most reliable method for diagnosing cancer [11]. Thus, analyzing the morphology and

spatial distribution of nuclei is essential for identifying cell types, assessing tissue structure, and determining the cancer category, which are crucial for diagnosing, evaluating severity, and planning treatment [26]. While traditional manual or semi-automated methods largely depended on the pathologist's expertise, in recent times, there have been continuous research and advancement in fully automated segmentation solutions offered by deep learning [14]. U-Net architecture [21] represents a benchmark in medical image segmentation tasks, where it utilizes symmetrical encoder-decoder based architecture for precise localization and skip connections for retaining contextual information. Numerous variations and improvements to the U-Net structures [29,19,20,6,4] were released that focused either on enhancing different aspects of the network or on adapting features from the datasets efficiently. Such approaches made U-Net being regarded as a task-specific model that performed well on a particular given set of task or datasets. Task-specific models are adept at capturing fine-grained details crucial for accurate cell segmentation but often require extensive training data and pre-processing, such as stain normalization, to handle the variability in staining and morphology characteristic of histopathological images. Furthermore, there is a need to comprehend broader context of tissue structures which is essential in understanding cell structure and formation of complex tissue architectures [22]. To overcome these challenges, there has been various research in leveraging the capabilities of foundational models, which are designed to understand and interpret a wide range of visual contexts [5,2,7]. Unlike task-specific models, foundation models such as the Segment Anything Model (SAM) [12] are trained on vast and varied datasets, enabling them to develop a more holistic understanding of images. While these models excel in identifying global features and contextual relationships, their application to histopathological images is limited by their inability to generate the detailed and fine segmentation necessary for accurate pathological analysis [22].

To address this gap, we propose leveraging the precision of U-Net architecture for local feature extraction and detailed segmentation while incorporating SAM for guiding this segmentation by providing broad, contextual insights to enhance the model's ability to understand and interpret complex, varied backgrounds and structures. In this paper, we propose - (1) enhancement to U-Net3+ [10] via adaptive feature selection which we call - *e*U-Net3+, and (2) X-Gated Fusion Block (X-GFB) to dynamically fuse the global and local latent representations. Our experimentation result shows an increase in performance of about 12% and 17.22% in CryoNuseg dataset [18]; about 15.55% and 16.77% in NuInsSeg dataset [17]; and about 9% improvements across dice scores and mIoUs using CoNIC dataset [8] compared to the base U-Net3+.

2 Methods

The overall architectural pipeline is shown in Fig. 1. Our proposed methodology first enhances U-Net3+ by adaptive feature selection for task-specific segmentation which we call *e*U-Net3+. Then we use frozen SAM encoder to guide the

segmentation process by providing global contextual features into the *e*U-Net3+. Both the local and global representations are then dynamically fused together using the proposed X-GFB, that first uses GLU in gated squeeze and excitation block [3] and then uses cross-attention block for retaining both local and global awareness.

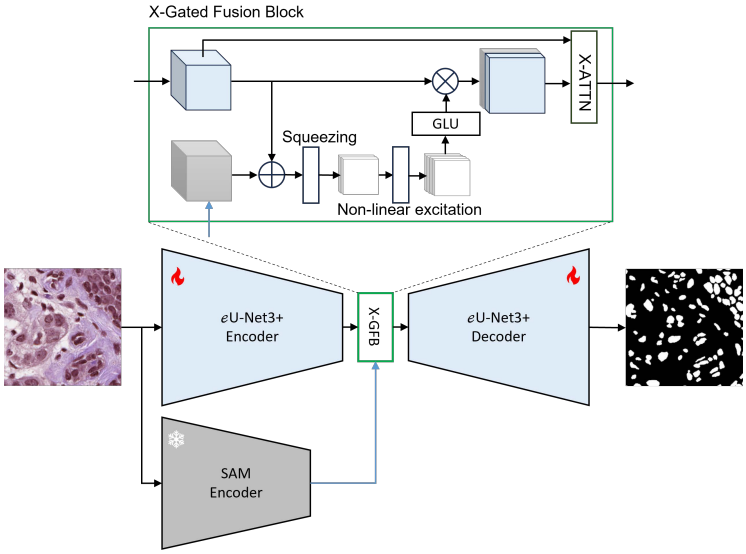

Fig. 1. The overall architecture of SAM guided task-specific segmentation.

2.1 *e*U-Net3+ as Task-Specific Model

The selection for the task-specific model necessitates the model to have the ability to capture both textural and structural information in great details from the image. The encoder-decoder based U-Net framework is a highly effective structure in the medical image segmentation models. It facilitates detailed feature extraction and critical integration for precise segmentation tasks [19]. Among the U-Net variants, U-Net3+ contains full-scale connections and deep supervisions that is aimed to capture multi-scale features more effectively. This approach enables the effective fusion of both high-level structural and low-level textural details from across the network and thus improving segmentation performance. This methodology ensures a comprehensive fusion of features, addressing the semantic gap between different network layers and enhancing the model's ability to produce refined semantic information [10].

Our implementation enhances U-Net3+ through incorporation of adaptive feature selection. Features can vary across different regions in a histopathological image. Certain regions might be densely-packed than others. Traditional

activation functions, such as ReLU [1], uniformly transform all features without discrimination, potentially overlooking the nuanced differences that are critical for accurate segmentation. To address this, we incorporate Gated Linear Units (GLUs) [23] into our U-Net3+ model for enabling selective feature activation based on their relevance to the segmentation task. GLUs operate by splitting the input into two streams: one undergoing a linear transformation and the other processed through a sigmoid activation function. The sigmoid's output effectively acts as a gate, modulating the flow of information from the linear stream based on the feature's importance. This is formulated as:

$$GLU(x) = sigmoid(W_a \cdot x + b_a) \odot (W_l \cdot x + b_l) \tag{1}$$

, where $\odot$ denotes the element wise multiplication, W_a, b_a denotes weights and biases for the gated mechanism, and W_l, b_l denotes the weights and biases for the linear transformation. This adaptive mechanism ensures that the eU-Net3+ model not only captures a broad spectrum of features from multi-scale inputs but also fine-tunes its focus towards the most diagnostically significant features.

2.2 SAM Encoder for Global Context

Segment Anything Model (SAM) is a promptable visual foundational model trained on one billion masks from 11M images for image segmentation [12]. There have been several research tasks that have used the zero-shot capabilities of SAM but struggled to produce high performing results (as can be seen in Fig. 2). This is due to the domain gap in the training images as the images used to train SAM were natural images which are very far from medical or histopathological images [16]. Even then, due its large training and generalizing capabilities, it can retain high-order global contextual information. We leverage this capability of SAM to guide our task-specific model towards increasing the segmentation performance. Upon comparing different model checkpoints from Fig. 2, it becomes evident that the the Huge model checkpoints (636M parameters) provides much better representation of the input image compared to the base model checkpoint (91M parameters). But this results in loss of ambiguity in certain regions as pointed out in label (f), (g) and (h) of Fig. 2. In our experiments, we wanted to provide more ambiguous global representations to eU-Net3+, so that the task specific model can decide on nuclei or non-nuclei regions as it has more fine-tuned understanding of feature characteristics. So, we chose the base model checkpoint with 91M parameters. The image encoding visualizations shown in Fig. 2 are plotted after performing PCA, where three most important components are picked from all the channels of the encoding and are treated as color channels [28].

2.3 X-Gated Fusion Block

Simple aggregation of feature embeddings from different encoders does not inherently guarantee enhanced performance due to potential conflicts or redundancies between the features [25]. We also confirm this visually in Fig. 3. To address

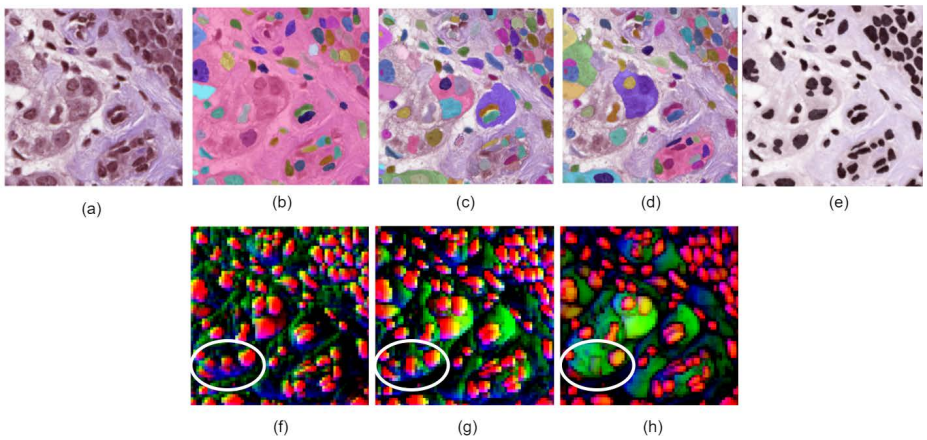

Fig. 2. (a) depicts a sample nuclei image and (e) depicts ground truth overlayed image. (b), (c), and (d) are the segmentation results predicted by SAM and (f), (g) and (h) visualization of PCA [28] for SAM encodings using SAM with Base, Large, and Huge model checkpoints, respectively. While the Huge model checkpoint yields a more distinct representation, it overlooks several ambiguous regions.

this issue, we design a dynamic fusion approach that can effectively integrate global context into the segmentation process and thereby improve the overall efficacy of the model. Inspired from the implementation of [3], where the authors used pose feature maps to guide the training of appearance feature maps for person re-identification, we use a similar Gated Squeeze and Excitation block followed by a Cross-Attention block to fuse the latent representations of SAM and eU-Net3+.

The squeeze part is performed using adaptive average pooling on the concatenated feature maps. This compresses the spatial dimensions and focuses on the global information contained within each channel. The excitation phase is then initiated by the convolution operation that expands the number of channels while preparing them for the gating mechanism implemented by the GLU, where it selectively allows information through the network. This ensures that the important features are dynamically selected forward. This selective gating mechanism ensures that only the most relevant features for segmentation are emphasized. The final convolution, followed by a sigmoid function, serves to recalibrate the channel-wise feature responses by learning nonlinear interactions between channels.

To ensure that the broader image context from SAM does not over-emphasis the local details of eU-Net3+, we implement a cross-attention block. The cross-attention mechanism operates on the flattened and permuted gated features, treating them as queries, keys, and values in the attention operation. This allows the model to amplify important features, suppress irrelevant information, and enhance contextual awareness. Fig. 3 provides a series of visualizations

that showcase the progressive transformation of image data as it passes through various layers of the *e*U-Net3+ architecture. By applying PCA to the high-dimensional feature space and projecting it down to the three principal components, we gain a visual insight of how the network distills information. Starting from (a) and moving through to (e), there is a noticeable transition from detailed, high-resolution features towards more abstract, aggregated representations as we move from shallower to deeper layers of the network. Although, representation of fifth layer (e) might be less recognizable in terms of original tissue morphology; however, they retain most essential characteristics. The subsequent image (g), which visualizes the concatenated features outputs of both SAM and *e*U-Net3+ encoder, can be observed to have similar representation as (e) with not much change, indicating that simple feature aggregation may not substantively effective. In contrast, (h) presents the transformed feature space after applying the proposed X-GFB. This visualization distinctly shows the integration of contextual information from SAM with the localized features of U-Net3+ using X-GFB to be more detailed compared to simple concatenation.

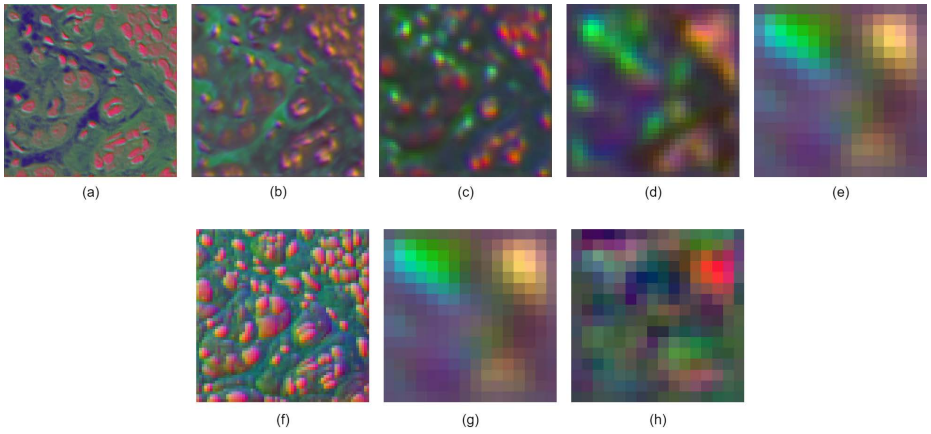

Fig. 3. Analysis of features after different layers and process in the network. (a) to (e) are visualizations of PCA that correspond to encodings from layer 1 to 5 of the *e*U-Net3+ encoder. (f) represents the SAM encoding, (g) showcase the feature space post-concatenation of SAM and *e*U-Net3+ encoder and (h) visualizes the feature space after applying X-GFB.

3 Experiments and Results

This section details the evaluation of the SAM Guided *e*U-Net3+ model across three histopathological datasets: CryoNuSeg, NuInsSeg, and CoNIC, chosen for their diversity in staining methods and tissue complexity. CryoNuSeg includes

a collection of 30 images, NuInsSeg contains 665 image patches, and CoNIC contains 4981 images in the dataset.

Our experiments are implemented using the PyTorch framework equipped with RTX A6000 graphics card. To standardize input, all images are resized to a uniform resolution of 256x256 pixels. The training process spans 50 epochs, with an initial learning rate set to 1×10^{-4} and batch size of 16. We use Adam optimization along with drop out of 0.3. For the loss function, we use a combination of weighted Dice [24] and focal loss [15] with equal weights, aiming to balance the training focus between prevalent and rare segmentation targets. Moreover, we use standard photo-metric and geometric augmentations to generate corresponding images for training the model. The models mentioned in Table 1 are trained and evaluated separately. We perform qualitative analysis as shown in Fig. 4, to assess the segmentation quality of with and without SAM guidance. *e*U-Net3+ is the enhanced U-Net3+ model with no SAM guidance seems to suffer in segmentation performance owing to the lack of global structural context.

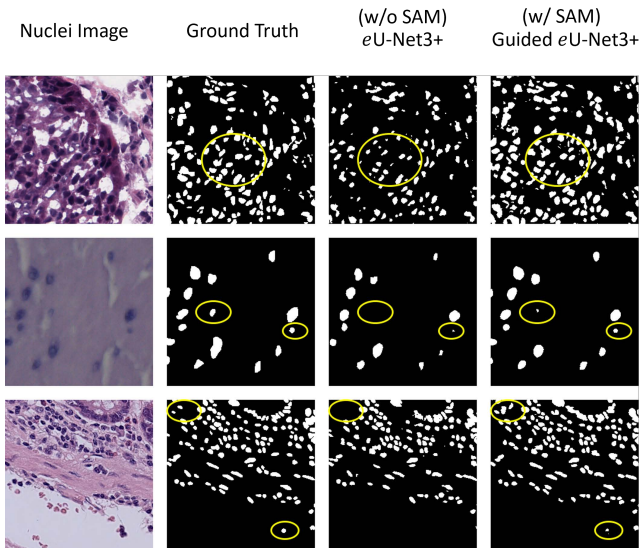

Fig. 4. Qualitative evaluation of segmentation performance of sample images for with and without SAM-guidance for *e*U-Net3+. The image in the top row is from CryoNuSeg, middle row is from NuInsSeg and bottom row is from CoNIC.

Fig. 4 shows the qualitative assessment our model. The circles areas in the image highlights some of the regions where inclusion of SAM in *e*U-Net3+ performed better than using only the task-specific model. The SAM guided predictions generally show more precise and accurate segmentation, further confirming the effects of of the proposal method of adding global context from SAM to improve the segmentation accuracy of the task-specific model.

Furthermore, table 1 summarizes the performance metrics quantitatively. The *e*U-Net3+ showed improvements with increases in dice scores of approximately 11.64%, 15.55%, and 8.77% for the CryoNuSeg, NuInsSeg, and CoNIC datasets, respectively. The SAM Guided *e*U-Net3+ model further increased the performance demonstrating superior segmentation over existing benchmarks. For *CryoNuSeg*, the model achieved a Dice score of 0.8942 and an mIoU of 0.8164; Dice score of 0.9399 and an mIoU of 0.8938 in *NuInsSeg* and Dice score of 0.9351 and an mIoU of 0.8869 on *CoNIC* across datasets, outperforming state-of-the-art task-specific models.

Table 1. Model Performance Across Different Datasets

Dataset	Model	Dice (F1-Score)	mIoU
CryoNuSeg	U-Net	0.7371	0.610
	DDU-Net [27]	0.8143	0.6822
	U-Net3+	0.778	0.6432
	*e*U-Net3+ (w/o SAM) (proposed)	0.8401	0.7644
	Guided *e*U-Net3+ (w/ SAM) (proposed)	**0.8942**	**0.8164**
NuInsSeg	U-Net	0.797	0.6781
	DDU-Net	0.7154	0.6133
	U-Net3+	0.7844	0.7261
	*e*U-Net3+ (w/o SAM) (proposed)	0.8307	0.8163
	Guided *e*U-Net3+ (w/ SAM) (proposed)	**0.9399**	**0.8938**
CoNIC	U-Net	0.7353	0.6214
	DDU-Net	0.827	0.7347
	U-Net3+	0.8474	0.7992
	*e*U-Net3+ (w/o SAM) (proposed)	0.8966	0.8539
	Guided *e*U-Net3+ (w/ SAM) (proposed)	**0.9351**	**0.8869**

4 Conclusion

In this study, we introduced a novel segmentation framework where we used SAM to provide the broad global representational information to the detailed local feature extraction task-specific *e*nhanced U-Net3+ model using X-GFB. Consequently, our model adeptly navigates the challenges inherent in histopathological image analysis, such as stain variability and complex tissue morphology. Our findings indicate that dynamically integrating global and local representations not only yields substantial improvements but also sets new benchmarks, outperforming existing state-of-the-art models in the field. Potential avenues for further enhancements include - exploring the integration of SAM with the hidden layers of U-Net3+ to provide additional structural context into the encoder and comparing the model performance against other SAM-adapted models. These areas, poised for future investigation, hold promise for advancing our understanding and capabilities in digital pathology and cancer diagnosis.

Acknowledgments. This research was supported by BL Science grants (202301630001).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu) (2019)
2. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D.: Foundational models in medical imaging: A comprehensive survey and future vision (2023)
3. Bhuiyan, A., Liu, Y., Siva, P., Javan, M., Ayed, I.B., Granger, E.: Pose guided gated fusion for person re-identification. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2664–2673 (2020). https://doi.org/10.1109/WACV45572.2020.9093370
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation (2021)
5. Chambon, P., Bluethgen, C., Langlotz, C.P., Chaudhari, A.: Adapting pretrained vision-language foundational models to medical imaging domains (2022)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021)
7. Gao, Y., Xia, W., Hu, D., Gao, X.: Desam: Decoupling segment anything model for generalizable medical image segmentation (2023)
8. Graham, S., Jahanifar, M., Vu, Q.D., Hadjigeorghiou, G., Leech, T., Snead, D., Raza, S.E.A., Minhas, F., Rajpoot, N.: Conic: Colon nuclei identification and counting challenge 2022 (2021)
9. Gross, S.M., Mohammadi, F., Sanchez-Aguila, C., Zhan, P.J., Liby, T.A., Dane, M.A., Meyer, A.S., Heiser, L.M.: Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. Nature Communications **14**(1), 3450 (Jun 2023). https://doi.org/10.1038/s41467-023-39122-z, https://doi.org/10.1038/s41467-023-39122-z
10. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation (2020)
11. Jiang, X., Hu, Z., Wang, S., Zhang, Y.: Deep learning for medical Image-Based cancer diagnosis. Cancers (Basel) **15**(14) (Jul 2023)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
13. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging **36**(7), 1550–1560 (2017). https://doi.org/10.1109/TMI.2017.2677499
14. Li, H., Zhong, J., Lin, L., Chen, Y., Shi, P.: Semi-supervised nuclei segmentation based on multi-edge features fusion attention network. PLoS One **18**(5), e0286161 (May 2023)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)

16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images (2023)
17. Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., Woitek, R., Hatamikia, S., Ellinger, I.: Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images (2023)
18. Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., Ellinger, I.: Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. Computers in Biology and Medicine **132**, 104349 (May 2021). https://doi.org/10.1016/j.compbiomed.2021.104349, http://dx.doi.org/10.1016/j.compbiomed.2021.104349
19. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (2018)
20. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (Oct 2020). https://doi.org/10.1016/j.patcog.2020.107404, http://dx.doi.org/10.1016/j.patcog.2020.107404
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
22. Ryu, J., Puche, A.V., Shin, J., Park, S., Brattoli, B., Lee, J., Jung, W., Cho, S.I., Paeng, K., Ock, C.Y., Yoo, D., Pereira, S.: Ocelot: Overlapped cell on tissue dataset for histopathology (2023)
23. Shazeer, N.: Glu variants improve transformer (2020)
24. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations, p. 240–248. Springer International Publishing (2017)
25. Wang, H., Vasu, P.K.A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., Pouransari, H.: Sam-clip: Merging vision foundation models towards semantic and spatial understanding (2023)
26. Wang, S., Rong, R., Zhou, Q., Yang, D.M., Zhang, X., Zhan, X., Bishop, J., Chi, Z., Wilhelm, C.J., Zhang, S., Pickering, C.R., Kris, M.G., Minna, J., Xie, Y., Xiao, G.: Deep learning of cell spatial organizations identifies clinically relevant insights in tissue images. Nature Communications **14**(1), 7872 (Dec 2023). https://doi.org/10.1038/s41467-023-43172-8, https://doi.org/10.1038/s41467-023-43172-8
27. Wang, Y., Peng, Y., Li, W., Alexandropoulos, G.C., Yu, J., Ge, D., Xiang, W.: Ddu-net: Dual-decoder-u-net for road extraction using high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2022). https://doi.org/10.1109/tgrs.2022.3197546
28. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 45533–45547. Curran Associates, Inc. (2023)
29. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation (2018)