Rethinking Decoder Design: Improving Biomarker Segmentation Using Depth-to-Space Restoration and Residual Linear Attention (Supplementary Material)

Saad Wazir[1] Daeyoung Kim[1]
saad.wazir@kaist.ac.kr, kimd@kaist.ac.kr
[1]School of Computing, KAIST, Republic of Korea

1. Supplementary Experimental Details

In this supplementary document, we provide additional insights into the dataset we used, extra results from our experiments, which were omitted from the paper due to page limitations. This document also includes experimental details and descriptions of evaluation metrics. We hope this supplementary information will assist the scientific community in understanding and replicating our research more effectively.

1.1. Datasets

Table 1 provides an overview of the datasets. The following paragraphs provide more specific information about each dataset.

1. **The Multi-organ Nucleus Segmentation (MoNuSeg)** [4] dataset contains H&E-stained histopathology images from 30 patients with tumors of the liver, kidney, prostate, bladder, breast, colon, and stomach organs, captured at 40 magnification. The dataset comprises 44 images, each with dimensions of 1000 by 1000 pixels containing 29,000 nuclear boundary annotations. For the purpose of training and evaluation, the MoNuSeg dataset consists of 30 images allocated for training and an additional 14 images reserved for testing.

2. **The 2018 Data Science Bowl (DSB)** [2] dataset encompasses a diverse collection of segmented nuclei images acquired under various conditions and from different organisms. The primary objective of this dataset is to challenge the generalization capabilities of models across these diverse variations. It includes 670 nuclei images, each accompanied by a segmented mask that corresponds to a single nucleus, with strict non-overlapping criteria between masks. The images have dimensions ranging from 128 x 128 to 512 × 512 pixels. The DSB dataset is bifurcated into two stages for training and testing purposes. In the first stage, there are 670 nuclei-segmented images and masks utilized for training, along with an additional 65 annotated images dedicated to testing. Subsequently, the second stage comprises an exclusive set of 3019 images meant solely for testing. In the development of our model, we solely relied on the stage 1 training set due to the unavailability of publicly accessible ground truth masks for stage 1 and stage 2 testing.

Table 1. Overview of datasets and their distribution.

Dataset	Biomarker	# Images - Train / Test	Classes	Image Size	After Offline Augmentation - Train
MoNuSeg	Nuclei	30 / 14	2	1000x1000	8820
DSB	Nuclei	603 / 67	2	128x128 - 512x512	13710
EM	Mitochondria	165 / 165	2	768x1024	9852
TNBC	Nuclei	- / 50	2	512x512	-

3. **Electron Microscopy (EM)** [5] dataset features annotated mitochondria from the CA1 hippocampus region of the brain, corresponding to a 1065x2048x1536 volume. This dataset is divided into two sub-volumes, each containing the first 165 slices of the image stack. The training and testing sets each contain 165 images with corresponding masks, with a size of 768x1024 pixels. Each voxel has a resolution of approximately 5x5x5 nm, and the data is provided in a multipage TIF format.

4. **The Triple-negative breast cancer (TNBC)** [6] dataset contains 50 H&E-stained breast histopathology images and their corresponding masks, each of dimension 512 × 512 pixels. There are 4022 annotated cell nuclei in the dataset. We use the TNBC dataset only for testing to evaluate the generalization capacity of the network.

1.2. Data Augmentation

We use Albumentations [1] and TensorFlow built-in ImageDataGenerator for online augmentations. The details of

Table 2. Summary of Data Augmentations

	Type	Argument / Value		Type	Argument / Value
Offline	RandomBrightnessContrast	p=1.0	**Online**	rotation_range	90
	GridDistortion	p=1.0		width_shift_range	0.3
	Transpose	p=1.0		height_shift_range	0.3
	ElasticTransform	$p = 1, alpha = 120, sigma = 120 * 0.05, alpha\_affine = 120 * 0.03$		shear_range	0.5
	RandomCrop	p=1.0		zoom_range	0.3
	-	-		horizontal_flip	True
	-	-		vertical_flip	True

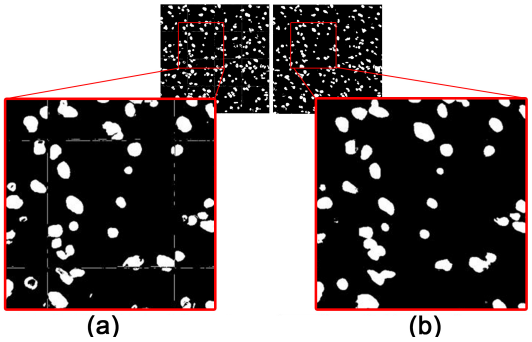

Figure 1. (a) Prediction with non overlapping patches IoU = 64.3% (b) Prediction with overlapping patches IoU = 65%

the data augmentations are shown in Table 2. For testing the overlapping patch technique was chosen to avoid the checkerboard effect seen with non-overlapping patches in some cases, as demonstrated in Figure 1.

1.3. Additional Qualitative Results

We provide additional qualitative results in Figure 2, to demonstrate the performance of each model on a diverse range of samples. We selected good, average, and bad samples based on an IoU threshold of 0.5 from models trained on the dataset separately. We highlight the over-segmented or falsely segmented areas with a red box. As you can observe, our model performs comparatively better under all diverse conditions.

1.4. Evaluation Metrics

To evaluate state-of-the-art deep learning methods and our proposed work, we have used standard evaluation metrics [3]. All the evaluation metrics are reported by calculating each evaluation metric for each prediction and taking the average over all samples in the test set. Following evaluation metrics are used:

1. **Intersection over Union (IoU)** [4] is defined as

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{1}$$

Where TP are True Positives , FP are False Positives, and FN are False Negatives. Intersection = TP, and Union = TP + FP + FN.

2. **Dice Coefficient (DSC)** [7] is defined as

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{2}$$

3. **Precision (Prec.)** is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

4. **Recall (Rec.)** is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

5. **False Omission Rate (FOR)** in binary semantic segmentation is the ratio of false negative predictions to the total number of negative predictions. This metric indicates the likelihood that a pixel predicted as negative by the model actually belongs to the segment. A higher false omission rate suggests the model may be missing relevant segments, leading to under-segmentation, while a lower FOR implies better accuracy in identifying true negatives. The False Omission Rate is defined as:

$$\text{False Omission Rate} = \frac{FN}{FN + TN} \tag{5}$$

6. **95th Percentile Hausdorff Distance (HD95)** is a metric used to measure the similarity between two sets of points, typically in the context of segmentation. It computes the 95th percentile of the distances from each point in one set to its nearest point in the other set. This measure helps mitigate the impact of outliers or noise in the segmentation. A lower HD95 indicates a closer match between the predicted and ground truth segmentations, suggesting better model performance, while a higher HD95 might point to greater discrepancies or misalignments. HD95 is defined as:

$$\begin{aligned} HD95_{AB} &= \text{percentile}_{95}\left(\min_{b \in B} d(a,b)\right), \forall a \in A \\ HD95_{BA} &= \text{percentile}_{95}\left(\min_{a \in A} d(b,a)\right), \forall b \in B \\ HD95 &= \max(HD95_{AB}, HD95_{BA}) \end{aligned} \tag{6}$$

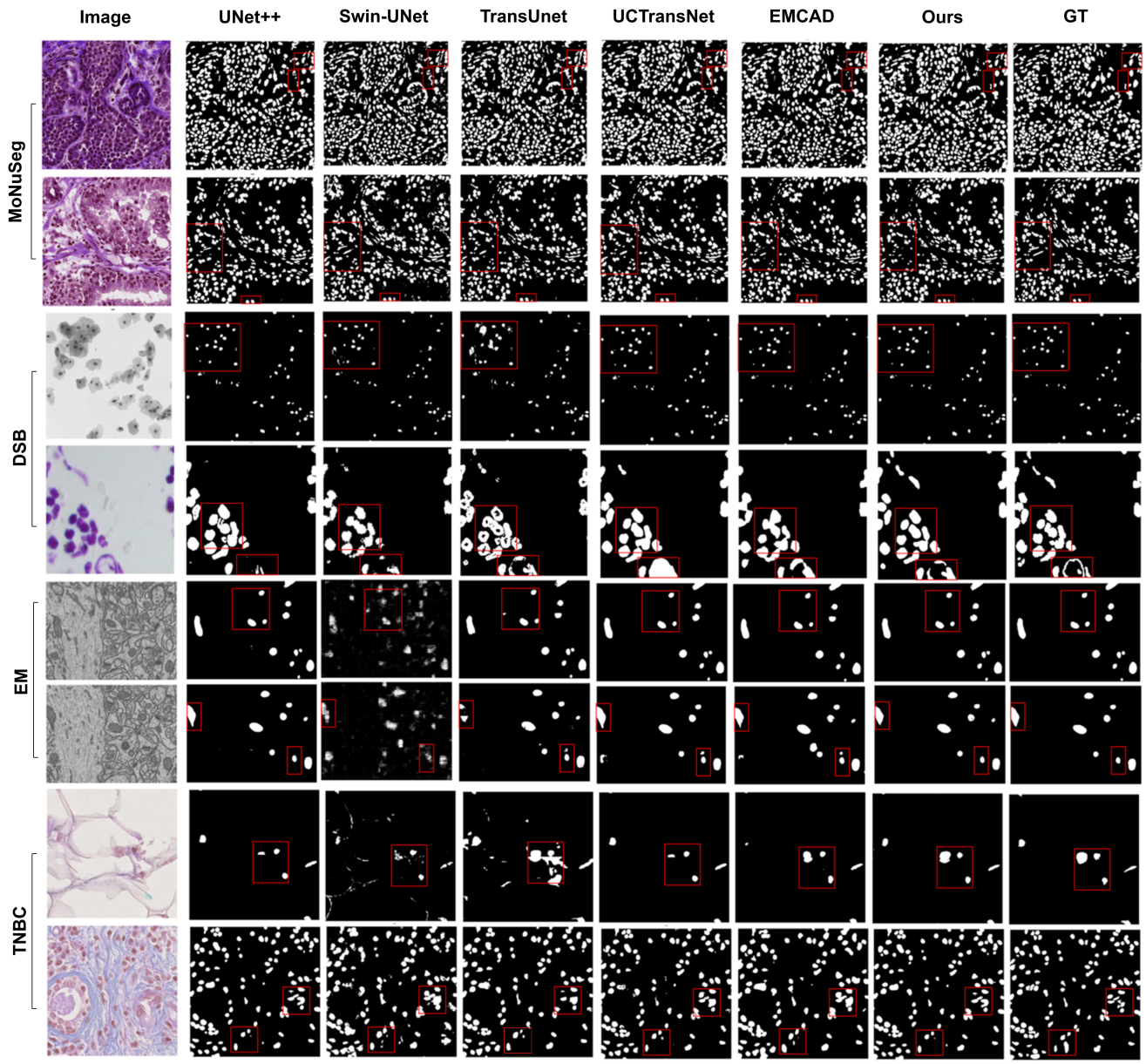

Figure 2. Additional Qualitative Results Comparison: Black pixels represent the background, while white pixels represent the biomarker. The red box indicates the region where there is either no prediction or over-segmentation.

Where: d(a, b) represents the Euclidean distance between two points a and b. The 'percentile_95' function returns the 95th percentile of the distances. The 'max' function computes the larger value among HD95_AB and HD95_BA.

7. **Average Surface Distance (ASD)** is a metric used in binary semantic segmentation to measure the average distance between the surfaces of predicted and ground truth segments. This metric assesses how well the model's predicted object boundaries align with those of the ground truth, focusing on the geometric accuracy of the segmentation. A lower ASD value indicates a closer match between the predicted and ground truth surfaces, suggesting a more accurate delineation of object boundaries. A higher ASD could suggest greater discrepancies, indicating potential misalignments or less accurate

segmentation. ASD is mathematically defined as:

$$\text{ASD} = \frac{1}{|A| + |B|} \left(\sum_{a \in A} \min_{b \in B} d(a, b) + \sum_{b \in B} \min_{a \in A} d(b, a) \right) \tag{7}$$

Where: A represents the set of points on the surface of the predicted segments. B represents the set of points on the surface of the ground truth segments. $d(a, b)$ represents the Euclidean distance between two points a and b.

References

[1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 1

[2] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 1

[3] Philip Meyer Florian Auer Iñaki Soto-Rey Frank Kramer Dominik Müller, Dennis Hartmann. Miseval: a metric library for medical image segmentation evaluation. *Studies in health technology and informatics*, 294:33–37, 2022. 2

[4] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36 (7):1550–1560, 2017. 1, 2

[5] Aurelien Lucchi, Yunpeng Li, and Pascal Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[6] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2):448–459, 2018. 1

[7] Thorvald Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948. 2