Multi-Class Cell Detection Using Modified Self-Attention

Tatsuhiko Sugimoto
Kyushu University, Japan
tatsu19962016@gmail.com

Hiroaki Ito
Kyoto University Hospital, Japan
hr27ito@kuhp.kyoto-u.ac.jp

Yuki Teramoto
Kyoto University Hospital
era1980@kuhp.kyoto-u.ac.jp

Akihiko Yoshizawa
Kyoto University Hospital
akyoshi@kuhp.kyoto-u.ac.jp

Ryoma Bise
Kyushu University, Japan
bise@ait.kyushu-u.ac.jp

Abstract

Multi-class cell detection (cancer or non-cancer) from a whole slide image (WSI) is an important task for pathological diagnosis. Cancer and non-cancer cells often have a similar appearance, so it is difficult even for experts to classify a cell from a patch image of individual cells. They usually identify the cell type not only on the basis of the appearance of a single cell but also on the context of the surrounding cells. For using such information, we propose a multi-class cell-detection method that introduces a modified self-attention to aggregate the surrounding image features of both classes. Experimental results demonstrate the effectiveness of the proposed method; our method achieved the best performance compared with a method, which simply uses the standard self-attention method.

1. Introduction

Cancer cell detection from a whole slide image (WSI) is an important task for pathological diagnosis. In some diagnoses, the rate of cancer cells over all the cells (cancer and non-cancer cells) is measured. For example, to check if cancer immunotherapy will be useful for a patient, a PDL1-test is widely used for diagnosis [3]. With this diagnosis, if the rate of cancer cells that are stained as brown over all the cancer cells is high, the therapy will be useful. To compute the rate, cancer cells must be counted from a WSI, which contains both cancer and non-cancer cells. However, it is time-consuming to count all the cells in a WSI because a WSI contains thousands of cells. In real diagnoses, pathologists roughly estimate the rates subjectively, and thus, their diagnoses often differ from amongst each other. Therefore, an automatic cancer cell counting system is required.

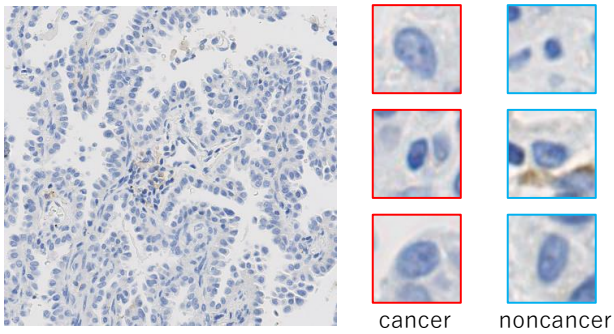


Figure 1. Examples of cancer and non-cancer cells. Left: wide-field view; Right: enlarged images of individual cells in each class. Some cells have similar appearances, and thus, it is difficult even for pathologists to identify class of cell from patch image. They usually use surrounding context of cell of interest.

To achieve this goal, it is necessary to detect all cells and classify these cells as cancer or non-cancer. The classification of cancer and non-cancer cells is challenging due to their similar appearances as shown in Fig. 1. The third rows of the enlarged images have very similar appearances. It is difficult even for pathologists to classify cells from only the enlarged patch images that contain individual cells. They identify the cell type not only on the basis of the appearance of a single cell but also on the context from the surrounding cells.

Cell detection for a single cell type has been studied in past decades, and high detection performances have been achieved. However, multi-class cell detection (such as cancer and non-cancer cells) is still challenging. Object detection for general objects, such as Faster R-CNN [23], can be directly applied to the multi-class cell detection task [17], which estimates bounding boxes of each cell and their class. This method

classifies the object in a bounding box on the basis of the local features. Therefore, it is difficult to accurately classify cells due to a lack of context from the surrounding cells. Other methods are based on Yolo v.3 [14], which performs multi-task learning, bounding box estimation and rough semantic-segmentation (i.e., a grid pixel belongs to a class). Rough semantic segmentation can use global spatial information unlike R-CNN based methods. The multi-class cell-detection method proposed by Abousamra et al. [1] simultaneously estimates the cell position mask (which contains all cell types) and the semantic segmentation masks for each class. It can use the spatial context for classifying each cell better compared to bounding box-based methods. However, it depends on the representation ability of the cell classifier, i.e., a single branch network has less feature extraction ability to classify cells.

In this paper, we propose a multi-class cell-detection method that has two decoders for estimating the cell position heat maps for each class, that is, cancer or non-cancer. To use the spatial context of the surrounding cells, we introduce a self-attention technique for aggregating image features of both classes, where self-attention has been used for transformers [10], such as ViT [11] and DETR [7]. Because the standard self-attention produces mixed features that contain both classes due to the summation operation, the network could not identify the discriminative features for cancer detection from the mixed feature. We thus propose a modified self-attention method that concatenates the features extracted by self-attention to the original features. The experimental results demonstrate the effectiveness of our method. The method using our modified self-attention improved the detection performance compared with that using the standard self-attention.

2. Related work

2.1. Cell detection

Cell detection has been studied in past decades, and it can be roughly categorized into three approaches: segmentation-based, bounding box-based, and point-based detection approaches. Segmentation methods can estimate detailed nuclei shapes. Many image-processing-based methods have been proposed for this approach: automatic threshold-based methods [4,21,25], watershed-based methods [8], and graph-cut-based methods [2]. However, when the cell population is dense, the methods often fail to segment the boundaries of cells that are touching because the boundaries are blurry, and small errors in the boundary segmentation cause mistakes, where several cells are identified as a single cluster. The deep learning-based method, such as convolutional neural network (CNN), has been popularly used for cell segmentation [15,16,19]. These methods outperform image-processing-based methods on datasets having various conditions. However, they require that the boundaries of cells be annotated, and thus, the annotation is time consuming.

In bounding box-based approach, general object detection methods, such as R-CNN [23], have been applied to cell detection tasks [17]. These bounding box-based methods classify cells on the basis of the local features in a bounding box. However, as discussed in the introduction, the context of the surrounding cells is important to identify the class of a cell.

Recently, point-based cell-detection methods have been proposed. One of the advantages of these methods is that the annotation cost is lower than the other two approaches: one click is only required for annotating one cell, and it is enough for cell counting. One major approach is estimating a cell position heat map, where a cell centroid position becomes a peak with a Gaussian distribution in the map [20]. This method can detect cells even in dense conditions, and it has shown the good detection performance in various tasks, such as weakly-supervised learning [20], domain adaptation [9], and learning from imperfect annotation [12]. However, these heat map-based methods have not been applied to multi-class cell detection.

Our method is categorized as a heat map-based method. If we simply use this method for multi-class detection, where the network has branches for each class, some of the detection results for each class may be duplicated (i.e., a cell belongs two classes) because neither of the decoders can use the discriminative features extracted from each other. To overcome this issue, we introduce a modified self-attention that can aggregate the features extracted from the two decoders.

2.2. Self-attention

Transformers have been successfully applied in many tasks, such as natural language processing [6, 10] and vision tasks [7, 11]. In the transformer for vision tasks, self-attention is used to aggregate a wide range of associative features among patch images, in which an image is separated into patches and these patches are fed into the transformer. These methods have been used for many vision tasks, such as classification [11], detection [7], and segmentation [13, 22]. They basically use self-attention for aggregating the spatial context among different patches. In contrast, our modified self-attention interchanges features extracted by different decoders for each class.

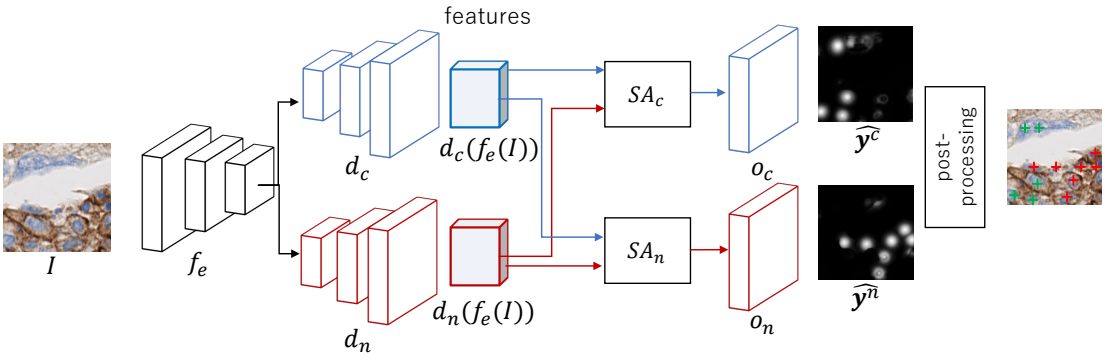


Figure 2. Overview of proposed method. Patch image is fed into network, and network then outputs cell positions for each class through two decoders and modified self-attention (mSA). mSA aggregates features of both classes, $d_c(f_e(I))$ for cancer and $d_c(f_e(I))$ for non-cancer. This framework can use context from surrounding cells to identify class of individual cells.

3. Multi-class cell detection by modified self-attention

3.1. Overview

Fig. 2 shows an overview of the proposed method. Given an image $\boldsymbol{I}$, the proposed method estimates two cell position heat maps: one for cancer cells $\hat{\boldsymbol{y}}^c$ and another for non-cancer cells $\hat{\boldsymbol{y}}^n$. The coordinates of the peaks in the maps indicates the centroid positions of cells for each cell type in the input image.

To estimate these heat maps, the network consists of seven sub-networks: 1) an encoder f_e for extracting features for both cell types, 2) and 3) decoders for extracting features of cancer cells d_c and non-cancer cells d_n, respectively, 4) and 5) modified self-attention modules SA_c, SA_n for aggregating the features of both types, 6) and 7) output layers (1×1 convolution) o_c, o_n for the results of detecting cancer and non-cancer, respectively. As discussed in the introduction, our key idea is that features of a different cell type are useful for detecting another type of cell. Therefore, we propose a modified self-attention module for aggregating the features of both cell types. In addition, we introduce post-processing to avoid duplicate detection results, which selects either cell type on the basis of the strength of the estimated signal in the same regions.

3.2. Cell position heat maps of each cell type

Many object detection methods use bounding boxes as the ground truth for object locations. Because the annotation of the bounding box is time consuming, we use point-level annotations, where a cell is annotated by one-click annotation (one click around the centroid of a cell). For point-level annotations, our method estimates cell-position heat maps for each class [20], which has been widely used and it has shown good performance.

Given a set of the annotated cell positions for an input image $\boldsymbol{I}$, the ground-truths of a cell-position heatmap $\boldsymbol{y}^c$ for cancer cell detection and $\boldsymbol{y}^n$ for non-cancer cell detection are generated so that the cell centroid position becomes a peak with a Gaussian distribution in the map. We generate an individual Gaussian distribution $\boldsymbol{y}_i^c$:

$$\boldsymbol{y}_i^c(\boldsymbol{u}) = \exp\left(-\frac{||\boldsymbol{u} - \boldsymbol{p}_i^c||_2^2}{\sigma^2}\right), \tag{1}$$

where i is an index of each cell, $\boldsymbol{u}$ indicates the position coordinate in the map, $\boldsymbol{p}_i^c$ indicates the ground-truth position of the i-th cancer cell, and σ adjusts the Gaussian blur as a hyper-parameter. Then, a cancer cell position heat map $\boldsymbol{y}^c$ is generated as the ground truth by maximizing $\boldsymbol{y}_i^c$ for each cell position $\boldsymbol{p}_i^c$:

$$\boldsymbol{y}^c = \max_i \boldsymbol{y}_i^c, \tag{2}$$

where the maximum operation is performed for each pixel, i.e., a pixel $\boldsymbol{u}$ takes a maximum value among $\boldsymbol{y}_i^c(\boldsymbol{u}), \forall i$. A non-cancer cell position heat map $\boldsymbol{y}^n$ is also generated in the same manner as $\boldsymbol{y}^c$ using the ground-truth position of non-cancer cells $\{\boldsymbol{p}_i^n\}_i$.

To train the entire network, we use the sum of the mean of the squared error (MSE) loss function between the predicted map $\hat{\boldsymbol{y}}$ and the ground-truth $\boldsymbol{y}$ for each cell type:

$$loss = \|\hat{\boldsymbol{y}}^c - \boldsymbol{y}^c\|^2 + \|\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n\|^2, \tag{3}$$

where both decoders are simultaneously trained with multi-task learning. The two decoders for each cell type have a representation ability that is better than a single network for classification, in which the segmentation results of each class are produced as channels.

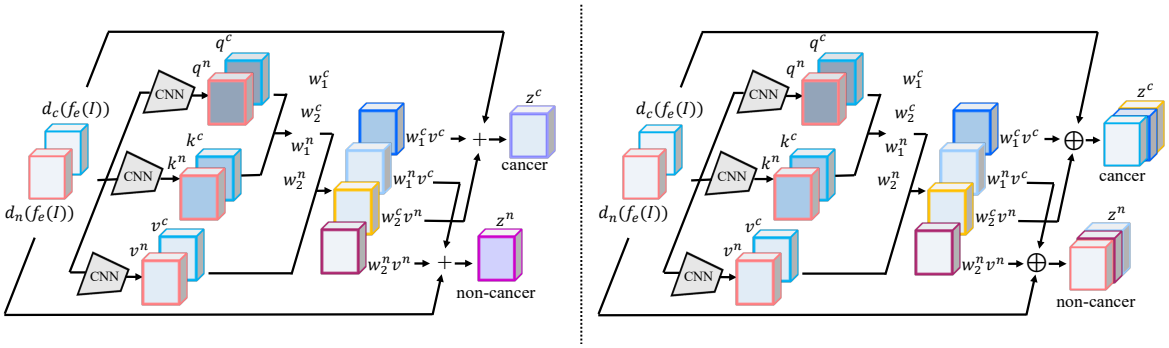


Figure 3. Illustration of standard self-attention (Left) and our modified self-attention (Right). w are computed by query and key (refer to Eq. 4 and 5), "+" is operation of summation, and $\oplus$ is operation of concatenation. Standard self-attention produces mixed features of both classes. In contrast, modified self-attention can use extracted features individually by concatenation.

3.3. Feature aggregation by modified self-attention

The network extracts discriminative features for each cell type. However, a cancer cell has a similar appearance to a non-cancer cell, and thus, the extracted features may still contain the features of both cell types (i.e., the network may not sufficiently disentangle these features on the basis of the local appearance). This has an adverse effect on cell detection even though post-processing is applied to avoid duplicate detection. As discussed in the introduction, to classify a cell, the context around the cell of interest is useful. Features for cancer cell detection do not contain information on the regions of surrounding cell of interest when the neighbor cells are non-cancer cells, i.e., the network cannot recognize that cells are sparsely placed (no neighbor cells) or that there are neighbor cells that are non-cancer cells, only from the features for cancer cell detection. Therefore, the features extracted by the other decoder for non-cancer cells are also useful for detecting cancer cells, and vice versa. For example, even if a cell appears to be similar to cancer cells, when cells around the cell of interest are estimated as non-cancer cells, this cell is more likely to be a non-cancer cell.

The self-attention technique has the potential to efficiently aggregate such information. However, features extracted by the standard self-attention contain the features of both cell types due to the summation operation of the original feature and the extracted features. This indicates that the output layer $\boldsymbol{o}_c$ cannot identify the discriminative features for cancer detection from mixed features. We thus propose a modified self-attention that concatenates the features extracted by self-attention to the original features.

Fig. 3 shows the summary of the difference of the standard and the modified self-attention module. We denote the input features for the modified self-attention module as $\boldsymbol{x}^c = d_c(f_e(\boldsymbol{I}))$ for cancer and $\boldsymbol{x}^n = d_n(f_e(\boldsymbol{I}))$ for non-cancer, which are outputs from decoders d_c and d_n. In the same manner as the standard self-attention [11], $\boldsymbol{x}^c$ is converted into a query $\boldsymbol{q}^c$, key $\boldsymbol{k}^c$, and value $\boldsymbol{v}^c$ by convolution, and the matrix is then flattened into a vector with dimension D. Using the features of the query and key, the weights of the features in the self-attention are defined as:

$$(w_1^c, w_2^c) = \mathrm{softmax}(\frac{1}{t}(\boldsymbol{q}^c \cdot \boldsymbol{k}^c, \boldsymbol{q}^c \cdot \boldsymbol{k}^n)), \quad (4)$$

$$(w_1^n, w_2^n) = \mathrm{softmax}(\frac{1}{t}(\boldsymbol{q}^n \cdot \boldsymbol{k}^c, \boldsymbol{q}^n \cdot \boldsymbol{k}^n)), \quad (5)$$

where w_1^c, w_2^c are the weights of features for cancer and w_1^n, w_2^n are weights for non-cancer, "$\cdot$" is an inner product operation, and t is a hyper-parameter for controlling the softmax operator. The weight w_1^c for the features of cancer cells becomes large when the query $\boldsymbol{q}^c$ of the features in the cancer cell detector is more similar to the key $\boldsymbol{k}^c$ in the cancer cell detector than the key $\boldsymbol{k}^n$ for non-cancer cell detection. On the basis of the estimated weights, the weighted sum of the values is concatenated to the original feature $\boldsymbol{x}^c$ as:

$$\boldsymbol{z}^c = \boldsymbol{x}^c \oplus w_1^c \boldsymbol{v}^c \oplus w_2^c \boldsymbol{v}^n, \quad (6)$$

$$\boldsymbol{z}^n = \boldsymbol{x}^n \oplus w_1^n \boldsymbol{v}^c \oplus w_2^n \boldsymbol{v}^n, \quad (7)$$

where $\oplus$ is the operation of concatenation, and $\boldsymbol{z}^c$ and $\boldsymbol{z}^n$ are the output features of the modified self attention module for cancer and non-cancer cells. Then, these extracted features are converted into cell position heat maps $\hat{\boldsymbol{y}}^c = o_c(\boldsymbol{z}^c)$, $\hat{\boldsymbol{y}}^n = o_n(\boldsymbol{z}^n)$ via 1×1 convolution layers o_c and o_n.

Using these concatenated features, the network can localize cells and identify the class of the cells. Let us consider a case where a non-cancer cell has a similar appearance to a cancer cell. In this case, the features

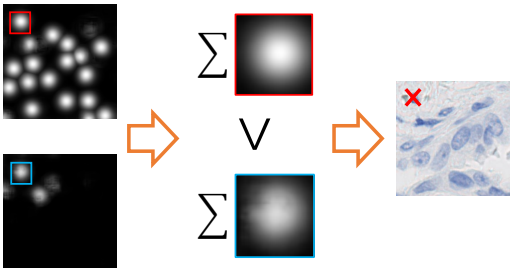


Figure 4. Overview of post-processing. If detection result is duplicated in both classes, sum of intensity in local patch is compared; cell belongs to class that has larger intensity than other class.

may be contained in both $\boldsymbol{v}_i^c$ and $\boldsymbol{v}_i^n$. If the neighbor cells are more likely to be cancer cells, $\boldsymbol{q}^c$ is similar to $\boldsymbol{k}^c$ than $\boldsymbol{k}^n$, and thus, the magnitude of the vector $w_1^c \boldsymbol{v}^c$ is larger than $w_2^c \boldsymbol{v}^n$. Using this information, the network can identify the cell of interest as cancer. If a cell is obviously a non-cancer cell, the features of the cell may be contained in only $\boldsymbol{v}_i^n$ and not in $\boldsymbol{v}_i^c$. In this case, even if the neighbor cells are cancer cells (i.e., w_1^c is large), the magnitudes of the region of the cell of interest $w_1^c \boldsymbol{v}_i^c$ are not large, and thus, the network can identify the cell as non-cancer.

3.4. Post-processing

The proposed modified self-attention module has the ability to avoid duplicate detection results by using the surrounding context of both classes. However, it is still not perfect, and duplicate detection remains (i.e., a cell could appear to be both cancer and non-cancer). To prevent the network from assigning multiple labels for one cell, we conduct post-processing in the inference step.

Fig. 4 shows the illustration of the post processing. We first find the duplicate detection points that are detected from both detectors for cancer and non-cancer cells, i.e., the detected cell has multiple labels (duplicate detection). To find the duplicate detection, we perform one-by-one matching for between detection results $\hat{\boldsymbol{y}}^c$ and $\hat{\boldsymbol{y}}^n$ on the basis of the distance of the detected points using linear programming. After matching, if the distance between the corresponding detected points is less than th_d, these detection points can be considered as duplicate detection.

For duplicate detection, we select the class on the basis of the heat maps $\hat{\boldsymbol{y}}^c$, $\hat{\boldsymbol{y}}^n$. We compare the sum of the intensity in the local area of the duplicate detection, in which the local area is a 64×64 bounding box in a 40x magnification image, whose center is the detected point in each map. If the value in the local area of $\hat{\boldsymbol{y}}^c$ is higher than that in $\hat{\boldsymbol{y}}^n$, the detected cell is a cancer cell, and vice versa.

4. Experiment

We evaluated our method using real data in the PD-L1 test. The PD-L1 test is one test for pathological diagnosis. The purpose of the PD-L1 test is to decide on a policy for cancer treatment. In this diagnosis, it is important to count the cancer cells from a WSI that contains both cancer and non-cancer cells. In real diagnosis, it is difficult to count all the cancer cells because a WSI contains thousands of cells. Therefore, they are usually roughly estimated subjectively. However, such subjective diagnoses tend to fluctuate among pathologists. Therefore, an automatic counting system is required.

4.1. Data set

We used a PD-L1 data set that was collected from a hospital. The data set contains 53 large images of different patients, in which the size of each image is about $9,000 \times 9,000$, and each image contains thousands of cells.

To make a ground truth, we conducted semi-automatic annotations using two types of annotations: cell localization by non-experts and cell classification by experts. It is easy even for non-experts to localize individual cells without classification. Therefore, the annotation for cell detection was performed by non-experts. Once given the ground-truth for cell detection, we trained a cell detection network and then applied the network to detect all the cells, which included cancer and non-cancer cells. From the results, most of the cells were well detected. However, it was difficult for non-experts to classify each cell. Thus, pathologists conducted annotation for classifying the detected cells. In this expert annotation, pathologists annotated regions that enclosed detected cells belonging to a single class (either cancer or non-cancer) and gave a class label for all the cells in the region at once. In this process, the pathologists confirmed the detected cells as correct.

As a result, the number of patch images was 20,092, which were cropped from large images of 53 patients with a size of 256×256 (40x magnifications). In the patches, the numbers of cancer cells and non-cancer cells were 16,703 and 72,084, respectively.

In the experiments, we split the data set into five sets for 5-fold cross-validation so that the different sets did not contain patch images cropped from the same patients. We used three sets for training, one for validation, and one for test in the 5-fold cross-validation in all the experiments.

4.2. Implementation details

The set of the encoder and decoder in our network has a U-net architecture [24], where the encoder and decoder were connected by skip connections. In our network, there are two decoders, thus one layer of the encoder has two skip connections for each decoder. The size of the output features from each decoder was $256 \times 256 \times 64$. In the modified self-attention module, we used convolutional layers for producing a key, query, and value. The output layers o^c and o^n consist of three layers of the 1×1 convolution, which estimate the heat maps of each class. We set $th_d = 20$, which was defined based on the cell size (a cell radius is about 20), and $t = 256 \times 8$, which was a default value in the transformer.

In training, we used Adam [18] for the optimization function with a learning rate of 10^{-5}. To stop the training, we used early stopping; if the validation loss had not improved during 10 epochs, we stopped training. Then, we selected the best model using the validation data.

4.3. Performance metric

For evaluation, we used three performance metrics: the mean of the recall (mRecall), the mean of the precision (mPrecision), and the F1 score as:

$$\begin{aligned} mRecall &= \frac{1}{M} \sum_{k \in \{c,n\}} \frac{TP_k}{TP_k + FN_k}, \\ mPrecision &= \frac{1}{M} \sum_{k \in \{c,n\}} \frac{TP_k}{TP_k + FP_k}, \\ F1 &= 2 \times \frac{mRecall \times mPrecision}{mRecall + mPrecision}, \end{aligned} \quad (8)$$

where mRecall and mPrecision are the mean of the precision of each class: cancer (c), and non-cancer (n). TP, FN, and FP are numbers of true positives, false negatives, and false positives, respectively. To define TP, FN, and FP, we conducted one-by-one matching among the detection results and the ground-truth in each class, which assigns each detection point to one ground-truth by minimizing the sum of the distances between corresponding positions using linear programming [5]. If the distance between a detected point and the assigned ground truth was less than a threshold (20 pixels) and had the same class label, we counted it as a TP. The threshold was determined on the basis of the cell size, where the radius of a cell is about 20 to 30 pixels. Non-assigned detection results were counted as FP and non-assigned ground-truth results were counted as FN.

Table 1. Performance metrics for each method.

	mPrecision	mRecall	F1
Nishimura	0.156	0.316	0.169
Multi task	0.855	0.687	0.759
Baseline	0.774	0.736	0.753
Proposed	0.870	0.739	**0.799**

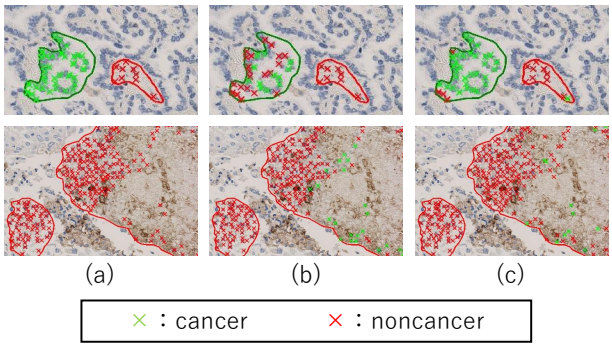
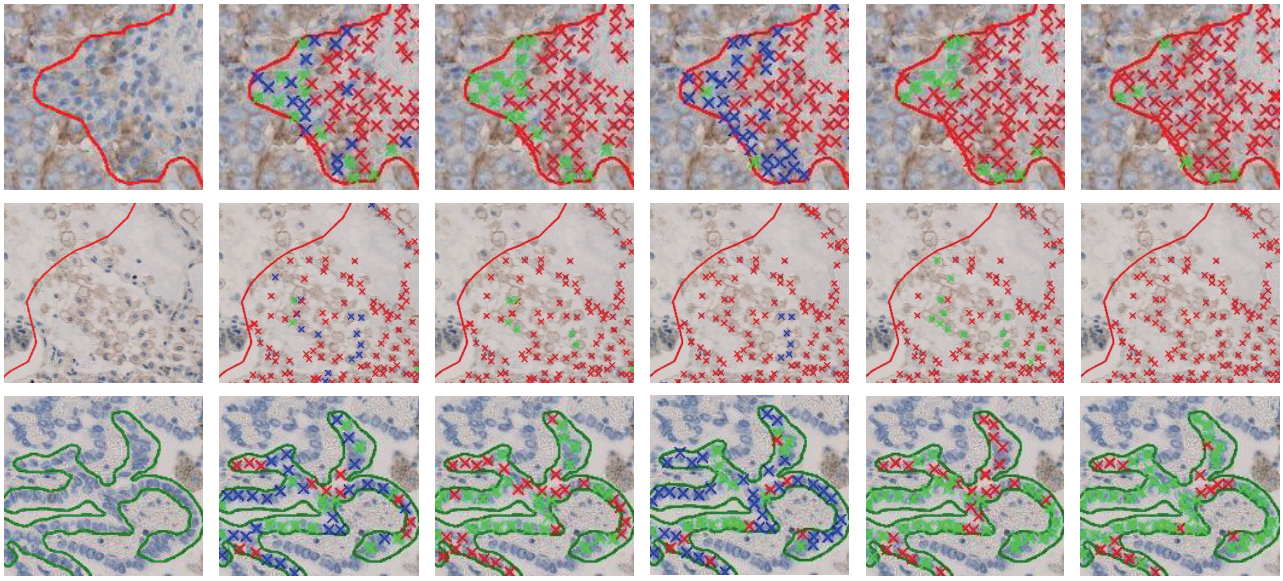


Figure 5. Examples of detection results for each method. (a)Ground truth, (b) multi-task learning-based method, and (c) ours. ×, ×indicate cancer and non-cancer cells. Areas enclosed by lines are annotated by experts but those outside are not.

4.4. Comparison with other methods

To confirm the effectiveness of our proposed network, we compared the proposed network with two current methods and the baseline method of ours: 1) Nishimura [20], which was designed for cell detection without classification, where the network estimates a cell position heat-map. Using this method, two U-nets for each class were trained individually; 2) Multi-task learning based on [1], which simultaneously performs detection for all cells and semantic segmentation for cell classed; 3) Baseline of our method, which was almost the same with our method except it did not use the self-attention mechanism and post-processing (i.e., it consisted of f_e, d_c, d_e, o_c, and o_n); and 4) our method, which uses the modified self-attention module.

Table. 1 shows the performance metrics of the methods. Nishimura's method, which trained the individual detection networks for each class, produced much overdetection, and thus, the performance was extremely worse since it is difficult to detect cancer cells from among many similar cells. In contrast, the baseline method improved the detection performance compared with individual training by sharing the features for detecting both classes. However, there was duplicate detection, which affected the performance. The multi-task learning-based method was better than the baseline method because the semantic-segmentation



(a) (b) (c) (d) (e) (f)

Figure 6. Examples of detection results in ablation study. (a) Original, (b) baseline, (c) with PP, (d) with modified SA without PP, (e) with standard SA , and (f) ours (modified SA). ×, ×denote cancer and non-cancer cells. ×denotes duplicate detection, which appears only in method without using post-processing.

Table 2. Ablation study

	post-process	standard SA	modified SA	mPrecision	mRecall	F1
Baseline				0.774	0.736	0.753
with PP	✓			0.848	0.711	0.773
with modified SA w/o PP			✓	0.791	0.761	0.775
with standard SA	✓	✓		0.838	0.739	0.785
ours (modified SA)	✓		✓	0.870	0.739	**0.799**

branch avoided the duplicate detection. Our method outperformed the multi-task learning-based method in terms of all metrics. We consider this to be because the modified self-attention module made it possible for the network to use the spatial context of the surrounding cells of both classes.

Fig. 5 shows examples of the detection results for each method, where the regions enclosed by lines were annotated by experts. Most of the cells were detected by all methods accurately. However, the classification performance was different among the methods. For multi-task learning, there were incorrect classification results in each class. Our method clearly improved the classification for both classes: cancer in Fig. 5 (Top) and non-cancer (Bottom).

4.5. Ablation study

Next, to show the effectiveness of each module in our method, we compared our method with four settings: a baseline that did not use the self-attention mechanism and post-processing (i.e., consisting of f_e, d_c, d_e, o_c, and o_n), a baseline with post-processing (PP) (i.e., consisting of f_e, d_c, d_e, o_c, o_n, and PP); a baseline with the modified self-attention without using post-processing; a baseline with the standard self-attention and post-processing. Our method used all modules.

Table 2 shows the detection results of each setting. When we used only the multi-decoder network for multi-class cell detection, the F1 score was worse. There were many instances of duplicate detection, and this had an adverse effect on the performance. Introducing post-processing improved the F1 score since there was no duplicate detection after the post-processing. The self-attention mechanism improved the mRecall and F1-score, but the mPrecision decreased. The standard self-attention method extracted the mixed features of cancer and non-cancer cells, and thus, the improvement was limited. Our method using the modified self-attention further improved all the metrics. Fig. 6 shows the detection re-

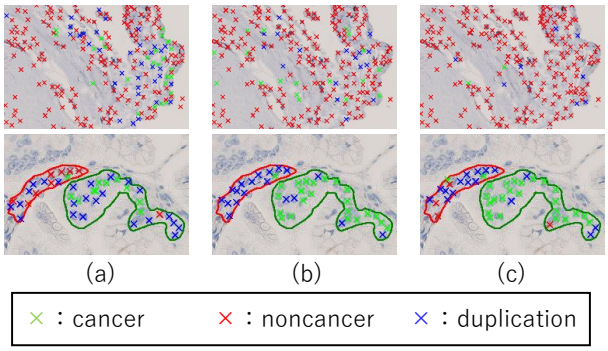


Figure 7. Examples of detection results. (a) Baseline (b) Baseline + standard self-attention, and (c) Baseline + modified self-attention. ×, ×denotes cancer and non-cancer cells. ×denote duplicate detection (i.e., cell was detected in heat maps of both classes).

Table 3. Evaluation of duplicate detection rate. "standard", "modified" indicate method used standard and modified self-attention, respectively.

	standard	modified	duplicate rate
Baseline			0.079
ours	✓		0.101
ours		✓	0.052

sults for each method, where the first and second rows show the examples of non-cancer cells, and the third row shows the examples of cancer cells. We can observe that our method clearly reduced the miss-classification results in both classes.

4.6. Effectiveness of modified self-attention for reducing duplicate detection

As discussed above, if the post-processing is not used, the method may produce duplicate detection results (i.e., a cell is detected in heat maps of both classes). We consider our modified self-attention module to have the ability to reduce duplicate detection results even without post-processing.

To show the effectiveness of the modified self-attention, we evaluated the rate of duplicate detection. It was defined as the number of instances of duplicate detection over the sum of the detection results in both classes. We compared the baseline method which has two decoders without interchanging features, the method that uses the standard self-attention module, and the proposed method, which uses the modified self-attention module for interchanging the features of each class. To evaluate the self-attention module itself for feature extraction, none of the methods used the post-processing.

Tab. 3 shows the duplication rate of each method. The baseline method had an 8 % duplicate detection rate. The method using the standard self-attention could not improve the duplicate detection even though it can improve the detection results as discussed above regarding the ablation study. Our modified self-attention module reduced duplicate detection from 0.079 to 0.052. This also had a good effect on overall cell detection results as shown in ablation study.

Fig. 7 shows examples of the detection results. The baseline method contained many duplicate detection results. As seen in the top row, the proposed method could greatly reduce duplicate detection. In the bottom row, our method reduced it in the green area, but duplicate detection still remained in the red area. We think that when cells that have very similar features are distributed densely, our method cannot improve the performance. However, when difficult cells are in discriminative cells as shown in Fig. 7 (Top) and the green region in (Bottom), our method could improve the detection performance.

5. Conclusion

In this paper, we proposed a multi-class cell detection method that estimates cell position heat maps for cancer cells and non-cancer cells. Because it is important to use the spatial context of the surrounding cells in addition to the features in the appearance of a cell of interest, we introduced a modified self-attention module for effectively using the image features of both classes. This can reduce duplicate detection by using the features extracted from different classes. As a result, the proposed method achieved better performance in comparison.

One of the limitations is that our method could not improve the detection performance in the case when cells that have very similar features are distributed densely because the aggregated features also contain ambiguity. We consider that the information aggregation by our method is still not sufficient to identify such cases. Since pathologists identify cancer cells using the global context, such as the spatial positional distribution, we will introduce a mechanism into the detection method to aggregate the further global spatial context and detailed features together in future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20H04211.

References

[1] Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee, Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, and Chao Chen. Multi-class cell detection using spatial context representation. In ICCV, pages 4005–4014, 2021. 2, 6

[2] Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. IEEE Trans. Biomed. Eng., 57(4):841–852, 2009. 2

[3] Hunter Ancevski, K.M.A. Socinski, and L.C. Villaruz. Pd-l1 testing in guiding patient selection for pd-1/pd-l1 inhibitor therapy in lung cancer. Mol Diagn Ther, 22, 2018. 1

[4] Ryoma Bise and Yoichi Sato. Cell detection from redundant candidate regions under nonoverlapping constraints. IEEE Trans. Med. Imag., 34(7):1417–1427, 2015. 2

[5] Ryoma Bise, Zhaozhen Yin, and Takeo Kanade. Reliable cell tracking by global data association. In ISBI, pages 1004–1010, 2011. 6

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In NeurIPS, volume 33, pages 1877–1901, 2020. 2

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, pages 213–229, 2020. 2

[8] Joe Chalfoun, Michael Majurski, Alden Dima, Christina Stuelten, Adele Peskin, and Mary Brady. Fogbank: a single cell segmentation across multiple cell lines and image modalities. Bmc Bioinformatics, 15(1):1–12, 2014. 2

[9] Hyeonwoo Cho, Kazuya Nishimura, Kazuhide Watanabe, and Ryoma Bise. Cell detection in domain shift problem using pseudo-cell-position heatmap. In MICCAI, 2021. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2018. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021. 2, 4

[12] Kazuma Fujii, Daiki Suehiro, Kazuya Nishimura, and Ryoma Bise. Cell detection from imperfect annotation by pseudo label selection using p-classification. In MICCAI, 2021. 2

[13] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In MICCAI, pages 61–71. Springer, 2021. 2

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014. 2

[15] Simon Graham, David Epstein, and Nasir Rajpoot. Dense steerable filter cnns for exploiting rotational symmetry in histology images. IEEE Trans. Med. Imag., 39(12):4124–4136, 2020. 2

[16] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal, 58:101563, 2019. 2

[17] Jane Hung, Allen Goodman, Deepali Ravel, Stefanie CP Lopes, Gabriel W Rangel, Odailton A Nery, Benoit Malleret, Francois Nosten, Marcus VG Lacerda, Marcelo U Ferreira, et al. Keras r-cnn: library for cell detection in biological images using deep neural networks. BMC bioinformatics, 21(1):1–7, 2020. 1, 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 6

[19] 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 Trans. Med. Imag., 36(7):1550–1560, 2017. 2

[20] Kazuya Nishimura, Ryoma Bise, et al. Weakly supervised cell instance segmentation by propagating from detection response. In MICCAI, 2019. 2, 3, 6

[21] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Trans. Syst., Man, Cybern., Syst., 9(1):62–66, 1979. 2

[22] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In MLMI, pages 267–276. Springer, 2021. 2

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS, 28, 2015. 1, 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234–241. Springer, 2015. 6

[25] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci. Transl. Med, 4(157):157ra143–157ra143, 2012. 2