HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images

Simon Graham[1,2,*], Quoc Dang Vu[3,*], Shan E Ahmed Raza[2,4], Ayesha Azam[2,5], Yee Wah Tsang[5], Jin Tae Kwak [3,+] and Nasir Rajpoot[2,6,+]

[*] *First authors contributed equally*
[+] *Last authors contributed equally*
[1] *Mathematics for Real World Systems Centre for Doctoral Training, University of Warwick, UK*
[2] *Department of Computer Science, University of Warwick, UK*
[3] *Department of Computer Science and Engineering, Sejong University, South Korea*
[4] *Centre for Evolution and Cancer & Division of Molecular Pathology, The Institute of Cancer Research, London, UK*
[5] *University Hospitals Coventry and Warwickshire, Coventry, UK*
[6] *The Alan Turing Institute, London, UK*

Abstract

Nuclear segmentation and classification within Haematoxylin & Eosin stained histology images is a fundamental prerequisite in the digital pathology workflow. The development of automated methods for nuclear segmentation and classification enables the quantitative analysis of tens of thousands of nuclei within a whole-slide pathology image, opening up possibilities of further analysis of large-scale nuclear morphometry. However, automated nuclear segmentation and classification is faced with a major challenge in that there are several different types of nuclei, some of them exhibiting large intra-class variability such as the nuclei of tumour cells. Additionally, some of the nuclei are often clustered together. To address these challenges, we present a novel convolutional neural network for simultaneous nuclear segmentation and classification that leverages the instance-rich information encoded within the vertical and horizontal distances of nuclear pixels to their centres of mass. These distances are then utilised to separate clustered nuclei, resulting in an accurate segmentation, particularly in areas with overlapping instances. Then, for each segmented instance the network predicts the type of nucleus via a devoted up-sampling branch. We demonstrate state-of-the-art performance compared to other methods on multiple independent multi-tissue histology image datasets. As part of this work, we introduce a new dataset of Haematoxylin & Eosin stained colorectal adenocarcinoma image tiles, containing 24,319 exhaustively annotated nuclei with associated class labels.

Keywords: Nuclear segmentation, nuclear classification, computational pathology, deep learning

1. Introduction

Current manual assessment of Haematoxylin and Eosin (H&E) stained histology slides suffers from low throughput and is naturally prone to intra- and inter-observer variability (Elmore et al., 2015). To overcome the difficulty in visual assessment of tissue slides, there is a growing interest in digital pathology (DP), where digitised whole-slide images (WSIs) are acquired from glass histology slides using a scanning device. This permits efficient processing, analysis and management of the tissue specimens (Madabhushi and Lee, 2016). Each WSI contains tens of thousands of nuclei of various types, which can be further analysed in a systematic manner and used for predicting clinical outcome. Here, the type of nucleus refers to the cell type in which it is located. For example, nuclear features can be used to predict survival (Alsubaie et al., 2018) and also for diagnosing the grade and type of disease (Lu et al., 2018). Also, efficient and accurate detection and segmentation of nuclei can facilitate good quality tissue segmentation (Sirinukunwattana et al., 2018; Javed et al., 2018), which can in turn not only facilitate the quantification of WSIs but may also serve as an important step in understanding how each tissue component contributes to disease. In order to use nuclear features for downstream analysis within computational pathology, nuclear segmentation must be carried out as an initial step. However, this remains a challenge because nuclei display a high level of heterogeneity and there is significant inter- and intra-instance variability in the shape, size and chromatin pattern between and within different cell types, disease types or even from one region to another within a single tissue sample. Tumour nuclei, in particular, tend to be present in clusters, which gives rise to many overlapping instances, providing a further challenge for automated segmentation, due to the difficulty of separating neighbouring instances.

As well as extracting each individual nucleus, determining the type of each nucleus can increase the diagnostic potential of current DP pipelines. For example, accurately classifying each nucleus to be from tumour or lymphocyte enables downstream analysis of tumour infiltrating lymphocytes (TILs), which have been shown to be predictive of cancer recurrence (Corredor et al., 2019). Yet, similar to nuclear segmentation, classifying the type of each nucleus is difficult, due to the high variance of nuclear appearance within each WSI. Typically, nuclei are classified using two disjoint models: one for detecting each nucleus and then another for performing nuclear classification (Sharma et al., 2015; Wang et al., 2016). However, it would be preferable to utilise a single unified model for nuclear instance segmentation and classification.

In this paper, we present a deep learning approach[1] for simultaneous segmentation and classification of nuclear instances in histology images. The network is based on the prediction of horizontal and vertical distances (and hence the name HoVer-Net) of nuclear pixels to their centres of mass, which are subsequently leveraged to separate clustered nuclei. For each segmented instance, the nuclear type is subsequently determined via a dedicated up-sampling branch. To the best of our knowledge, this is the first approach that achieves instance segmentation and classification within the same network. We present comparative results on six independent multi-tissue histology image datasets and demonstrate state-of-the-art performance compared to other recently proposed methods. The main contributions of this work are listed as follows:

- A novel network, targeted at simultaneous segmentation and classification of nuclei, where horizontal and vertical distance map predictions separate

[1] Model code available at: https://github.com/vqdang/hover_net

clustered nuclei.

- We show that the proposed HoVer-Net achieves state-of-the-art performance on multiple H&E histology image datasets, as compared to over a dozen recently published methods.
- An interpretable and reliable evaluation framework that effectively quantifies nuclear segmentation performance and overcomes the limitations of existing performance measures.
- A new dataset[2] of 24,319 exhaustively annotated nuclei within 41 colorectal adenocarcinoma image tiles.

2. Related Work

2.1. Nuclear Instance Segmentation

Within the current literature, **energy-based** methods, in particular the watershed algorithm, have been widely utilised to segment nuclear instances. For example, Yang et al. (2006) used thresholding to obtain the markers and the energy landscape as input for watershed to extract the nuclear instances. Nonetheless, thresholding relies on a consistent difference in intensity between the nuclei and background, which does not hold for more complex images and hence often produces unreliable results. Various approaches have tried to provide an improved marker for marker-controlled watershed. Cheng et al. (2009) used active contours to obtain the markers. Veta et al. (2013) used a series of morphological operations to generate the energy landscape. However, these methods rely on the predefined geometry of the nuclei to generate the markers, which determines the overall accuracy of each method. Notably, Ali and Madabhushi (2012) avoided the trouble of refining the markers for watershed

[2] The CoNSeP dataset for nuclear segmentation is available at `https://warwick.ac.uk/fac/sci/dcs/research/tia/data/`.

by designing a method that relies solely on the energy landscape. They combined an active contour approach with nuclear shape modelling via a level-set method to obtain the nuclear instances. Despite its widespread usage, obtaining sufficiently strong markers for watershed is a non-trivial task. Some methods have departed from the energy-based approach by utilising the geometry of the nuclei. For instance, Wienert et al. (2012), LaTorre et al. (2013) and Kwak et al. (2015) computed the concavity of nuclear clusters, while Liao et al. (2016) used eclipse-fitting to separate the clusters. However, this assumes a predefined shape, which does not encompass the natural diversity of the nuclei. In addition, these methods tend to be sensitive to the choice of manually selected parameters.

Recently, **deep learning** methods have received a surge of interest due to their superior performance in many computer vision tasks (Litjens et al., 2017; Shen et al., 2017; LeCun et al., 2015). These approaches are capable of automatically extracting a representative set of features, that strongly correlate with the task at hand. As a result, they are preferable to hand-crafted approaches, that rely on a selection of pre-defined features. Inspired by the Fully Convolutional Network (FCN) (Long et al., 2015), U-Net (Ronneberger et al., 2015) has been successfully applied to numerous segmentation tasks in medical image analysis. The network has an encoder-decoder design with skip connections to incorporate low-level information and uses a **weighted loss function** to assist separation of instances. However, it often struggles to split neighbouring instances and is highly sensitive to pre-defined parameters in the weighted loss function. A more recently proposed method in Micro-Net (Raza et al., 2018) extends U-Net by utilising an enhanced network architecture with weighted loss. The network processes the input at multiple resolutions and as a result, gains robustness against nuclei with varying size. In Graham and Rajpoot (2018), the authors developed a network that is robust to stain variations in H&E images by introducing a weighted loss function that is sensitive to the Haematoxylin intensity within the image.

Other methods exploit information about the nuclear **contour** (or boundary) within the network, such as DCAN (Chen et al., 2016) that utilised a dual architecture that outputs the nuclear cluster and the nuclear contour as two separate prediction maps. Instance segmentation is then achieved by subtracting the contour from the nuclear cluster prediction. Cui et al. (2018) proposed a network to predict the inner nuclear instance, the nuclear contour and the background. The network utilised a customised weighted loss function based on the relative position of pixels within the image to improve and stabilise the inner nuclei and contour prediction. Some other methods have also utilised the nuclear contour to achieve instance segmentation. For example, Kumar et al. (2017) employed a deep learning technique for labelling the nuclei and the contours, followed by a region growing approach to extract the final instances. Khoshdeli and Parvin (2018) used the contour predictions as input into a further network for segmentation refinement. Zhou et al. (2019) proposed CIA-Net, that utilises a multi-level information aggregation module between two task-specific decoders, where each decoder segments either the nuclei or the contours. A Deep Residual Aggregation Network (DRAN) was proposed by Vu et al. (2018) that uses a multi-scale strategy, incorporating both the nuclei and nuclear contours to accurately segment nuclei.

There have been various other methods to achieve instance separation. Instead of considering the contour, Naylor et al. (2018) proposed a deep learning approach to detect superior markers for watershed by regressing the nuclear **distance map**. Therefore, the network avoids making a prediction for areas with indistinct contours.

In line with these developments, the field of instance segmentation within natural images is also rapidly progressing and have had a significant influence on nuclear instance segmentation methods. A notable example is Mask-RCNN (He et al., 2017), where instance segmentation approach is achieved by first predicting candidate regions likely to contain an object and then deep learning based segmentation within those proposed regions.

2.2. Nuclear Classification

As well as performing instance segmentation, it is desirable to determine the *type* of each nucleus to facilitate and improve downstream analysis. It is possible for current models to differentiate between certain nuclear types in H&E, however sub-typing of lymphocytes is an extremely hard task due to the high levels of similarity in morphological appearance between T and B lymphocytes. Typically, classifying each nucleus is done via a two-stage approach, where the first step involves either nuclear segmentation or nuclear detection. When segmentation is used as the initial step, a series of morphological and textural features are extracted from each instance, which are then used within a classifier to determine the nuclei classes. For example, Nguyen et al. (2011) classified nuclei within H&E stained breast cancer images as either tumour, lymphocyte or stromal based on their morphological features. Yuan et al. (2012) performed nuclear segmentation and then classified each nucleus with AdaBoost classifier, utilising the intensity, morphology and texture of nuclei as features. Otherwise, detection is performed as an initial step and a patch centred at the point of detection is fed into a classifier, to predict the type of nucleus. Sirinukunwattana et al. (2016) proposed a spatially constrained CNN, that initially detects all nuclei and then for each nucleus an ensemble of associated patches are fed into a CNN to predict the type to be either epithelial, inflammatory, fibroblast or miscellaneous.

3. Methods

Our overall framework for automatic nuclear instance segmentation and classification can be observed in Fig. 1 and the proposed network in Fig. 2. Here, nuclear pixels are first detected and then, a tailored post-processing pipeline is used to simultaneously segment nuclear instances and obtain the corresponding nuclear types. The framework is based upon the horizontal and vertical distance maps, which can be seen in Fig. 3. In the figure, each nuclear pixel denotes either the horizontal or vertical distance of pixels to their centres of mass.

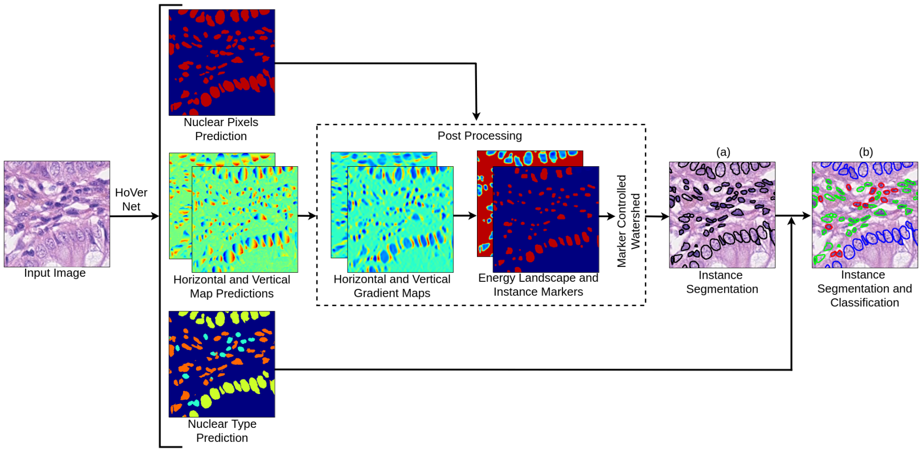


Figure 1: Overview of the proposed approach for simultaneous nuclear instance segmentation and classification. When no classification labels are available, the network produces the instance segmentation as shown in (a). The different colours of the nuclear boundaries represent different types of nuclei in (b).

3.1. Network Architecture

In order to extract a strong and representative set of features, we employ a deep neural network. The feature extraction component of the network is inspired by the pre-activated residual network with 50 layers (He et al., 2016) (Preact-ResNet50), due to its excellent performance in recent computer vision tasks (Deng et al., 2009) and robustness against input perturbation (Arnab et al., 2017). Compared to the standard Preact-ResNet50 implementation, we reduce the total down-sampling factor from 32 to 8 by using a stride of 1 in the first convolution and removing the subsequent max-pooling operation. This ensures that there is no immediate loss of information that is important for performing an accurate segmentation. Various residual units are applied throughout the network at different down-sampling levels. A series of consecutive residual units is denoted as a residual block. The number of residual units within each residual block is 3, 4, 6 and 3 that are applied at down-sampling levels 1, 2, 4 and 8 respectively. For clarity, a down-sampling level of 2 means that the input has a reduction in the spatial resolution by a factor of 2.

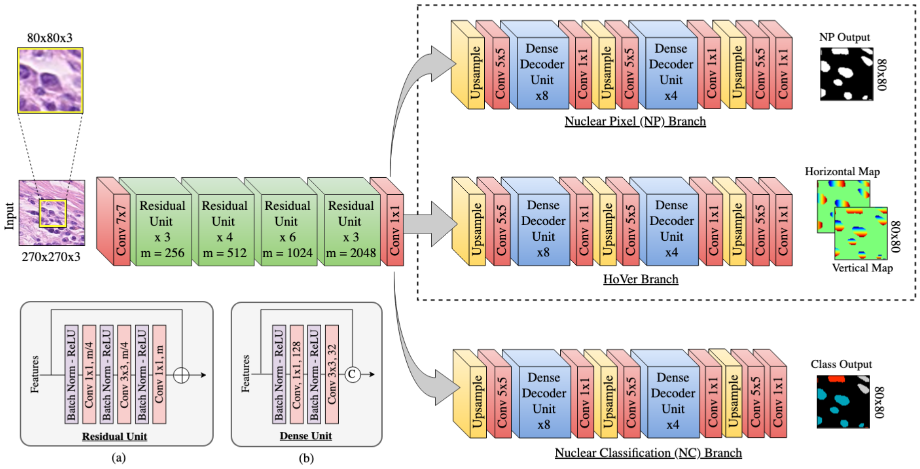


Figure 2: Overview of the proposed architecture. (a) (Pre-activated) residual unit, (b) dense unit. m indicates the number of feature maps within each residual unit. The yellow square within the input denotes the considered region at the output. When the classification labels aren't available, only the up-sampling branches in the dashed box are considered.

Following Preact-ResNet50, we perform nearest neighbour up-sampling via three distinct branches to simultaneously obtain accurate nuclear instance segmentation and classification. We name the corresponding branches: (i) nuclear pixel (NP) branch; (ii) HoVer branch and (iii) nuclear classification (NC) branch. The NP branch predicts whether or not a pixel belongs to the nuclei or background, whereas the HoVer branch predicts the horizontal and vertical distances of nuclear pixels to their centres of mass. Then, the NC branch predicts the type of nucleus for each pixel. In particular, the NP and HoVer branches jointly achieve nuclear instance segmentation by first separating nuclear pixels from the background (NP branch) and then separating touching nuclei (HoVer branch). The NC branch determines the type of each nucleus by aggregating the pixel-level nuclear type predictions within each instance.

All three up-sampling branches utilise the same architectural design, which consists of a series of up-sampling operations and densely connected units (Huang et al., 2016) (or dense units). By stacking multiple and relatively cheap dense units, we build a large receptive field with minimal parameters, compared to using a single convolution with a larger kernel size and we ensure efficient gradient propagation. We use skip connections (Ronneberger et al., 2015) to incorporate features from the encoder, but utilise summation as opposed to concatenation. The consideration of low-level information is particularly important in segmentation tasks, where we aim to precisely delineate the object boundaries. We use dense units after the first and second up-sampling operations, where the number of units is 4 and 8 respectively. Valid convolution is performed throughout the two up-sampling branches to prevent poor predictions at the boundary. This results in the size of the output being smaller than the size of the input. As opposed to using a dedicated network for each task, a shared encoder makes it possible to train the nuclear instance segmentation and classification model end-to-end and therefore, reduce the total training time. Furthermore, a shared encoder can also take advantage of the shared information across multiple tasks and thus, help to improve the model performance on all tasks.

Finally, if we do not have the classification labels of the nuclei, only the NP and HoVer up-sampling branches are considered. Otherwise, we consider all three up-sampling branches and perform simultaneous nuclear instance segmentation and classification.

We display an overview of the network architecture in Fig. 2, where the spatial dimension of the input is 270×270 and the output dimension of each branch is 80×80. The dashed box within Fig. 2 highlights the branches for nuclear instance segmentation. Additionally, we also show a residual unit and a dense unit within Fig. 2a and Fig. 2b. We denote m as the number of feature maps within each convolution of a given residual unit. At each down sampling level, from left to right, m=256, 512, 1024, 2048 respectively. We keep a fixed amount of feature maps within each dense unit throughout the two branches as shown in Fig. 2c.

3.1.1. Loss Function

The proposed network design has 4 different sets of weights: w_0, w_1, w_2 and w_3 which refer to the weights of the Preact-ResNet50 encoder, the HoVer

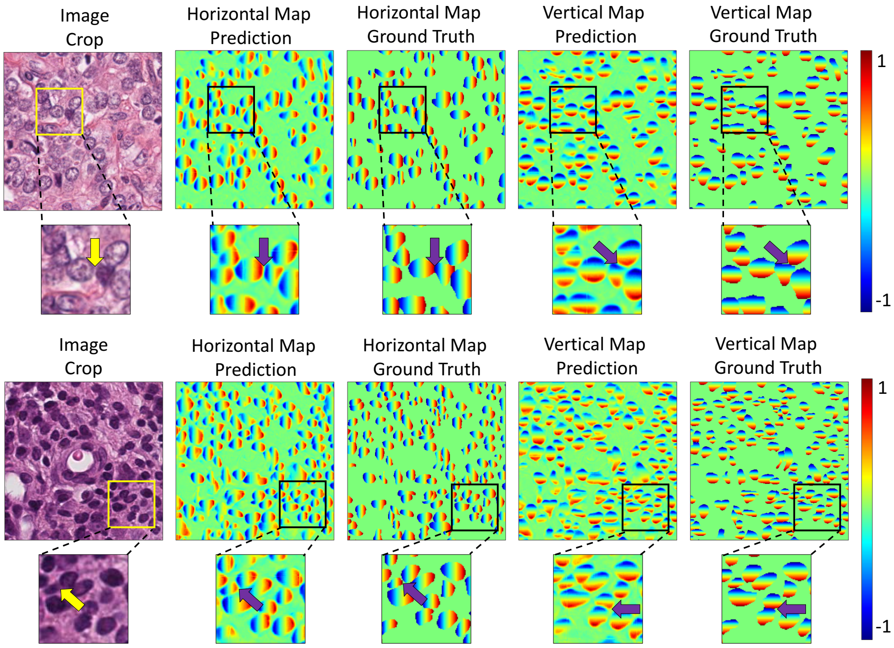


Image Crop
Horizontal Map Prediction
Horizontal Map Ground Truth
Vertical Map Prediction
Vertical Map Ground Truth
1
-1
Image Crop
Horizontal Map Prediction
Horizontal Map Ground Truth
Vertical Map Prediction
Vertical Map Ground Truth
1
-1

Figure 3: Cropped image regions showing horizontal and vertical map predictions, with corresponding ground truth. Arrows highlight the strong instance information encoded within these maps, where there is a significant difference in the pixel values.

branch decoder, the NP branch decoder and the NC branch decoder. These 4 sets of weights are optimised jointly using the loss $\mathcal{L}$ defined as:

$$\mathcal{L} = \underbrace{\lambda_a\mathcal{L}_a + \lambda_b\mathcal{L}_b}_{\text{HoVer Branch}} + \underbrace{\lambda_c\mathcal{L}_c + \lambda_d\mathcal{L}_d}_{\text{NP Branch}} + \underbrace{\lambda_e\mathcal{L}_e + \lambda_f\mathcal{L}_f}_{\text{NC Branch}} \tag{1}$$

where $\mathcal{L}_a$ and $\mathcal{L}_b$ represent the regression loss with respect to the output of the HoVer branch, $\mathcal{L}_c$ and $\mathcal{L}_d$ represent the loss with respect to the output at the NP branch and and finally, $\mathcal{L}_e$ and $\mathcal{L}_f$ represent the loss with respect to the output at the NC branch. We choose to use two different loss functions at the output of each branch for an overall superior performance. $\lambda_a ... \lambda_f$ are scalars that give weight to each associated loss function. Specifically, we set λ_b to 2 and the other scalars to 1, based on empirical selection.

Given the input image I, at each pixel i we define $p_i(I, w_0, w_1)$ as the regression output of the HoVer branch, whereas $q_i(I, w_0, w_2)$ and $r_i(I, w_0, w_3)$

denote the pixel-based softmax predictions of the NP and NC branches respectively. We also define $\mathit{\Gamma}_i(I)$, $\mathit{\Psi}_i(I)$ and $\mathit{\Phi}_i(I)$ as their corresponding ground truth (GT). $\mathit{\Psi}_i(I)$ is the GT of the nuclear binary map, where background pixels have the value of 0 and nuclear pixels have the value 1. On the other hand, $\mathit{\Phi}_i(I)$ is the nuclear type GT where background pixels have the value 0 and any integer value larger than 0 indicates the type of nucleus. Meanwhile, $\mathit{\Gamma}_i(I)$ denotes the GT of the horizontal and vertical distances of nuclear pixels to their corresponding centres of mass. For $\mathit{\Gamma}_i(I)$, we assign values between -1 and 1 to nuclear pixels in both the horizontal and vertical directions. We assign the value of the background and the line crossing the centre of mass within each nucleus to be 0. For clarity, we denote the horizontal and vertical components of the GT HoVer map as horizontal map $\mathit{\Gamma}_{i,x}$ and vertical map $\mathit{\Gamma}_{i,y}$ respectively. Visual examples of the horizontal and vertical maps can be seen in Fig. 3.

At the output of the HoVer branch, we compute a multiple term regression loss. We denote $\mathcal{L}_a$ as the mean squared error between the predicted horizontal and vertical distances and the GT. We also propose a novel loss function $\mathcal{L}_b$ that calculates the mean squared error between the horizontal and vertical gradients of the horizontal and vertical maps respectively and the corresponding gradients of the GT. We formally define $\mathcal{L}_a$ and $\mathcal{L}_b$ as:

$$\mathcal{L}_a = \frac{1}{n}\sum_{i=1}^{n}\left(p_i(I;\boldsymbol{w}_0,\boldsymbol{w}_1) - \mathit{\Gamma}_i(I)\right)^2 \tag{2}$$

$$\begin{aligned}\mathcal{L}_b = &\frac{1}{m}\sum_{i\in M}\left(\nabla_x(p_{i,x}(I;\boldsymbol{w}_0,\boldsymbol{w}_1)) - \nabla_x(\mathit{\Gamma}_{i,x}(I))\right)^2 \\ +&\frac{1}{m}\sum_{i\in M}\left(\nabla_y(p_{i,y}(I;\boldsymbol{w}_0,\boldsymbol{w}_1)) - \nabla_y(\mathit{\Gamma}_{i,y}(I))\right)^2\end{aligned} \tag{3}$$

Within equation (3), ∇_x and ∇_y denote the gradient in the horizontal x and vertical y directions respectively. m denotes total number of nuclear pixels within the image and M denotes the set containing all nuclear pixels.

At the output of NP and NC branches, we calculate the cross-entropy loss ($\mathcal{L}_c$ and $\mathcal{L}_e$) and the dice loss ($\mathcal{L}_d$ and $\mathcal{L}_f$). These two losses are then added together to give the overall loss of each branch. Concretely, we define the cross entropy and dice losses as:

$$\text{CE} = -\frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{K} X_{i,k}(I)\log Y_{i,k}(I) \tag{4}$$

$$\text{Dice} = 1 - \frac{2\times\sum_{i=1}^{N}(Y_i(I)\times X_i(I))+\epsilon}{\sum_{i=1}^{N}Y_i(I)+\sum_{i=1}^{N}X_i(I)+\epsilon} \tag{5}$$

where X is the ground truth, Y is the prediction, K is the number of classes and ϵ is a smoothness constant which we set to $1.0e^{-3}$. When calculating $\mathcal{L}_c$ and $\mathcal{L}_d$ for NP branch, for a given pixel i, we set X_i and Y_i as $q_i(I, w_0, w_2)$ and Ψ_i respectively. For $\mathcal{L}_c$, we set K to be 2 within equation (4) because the task of the branch is to perform binary nuclear segmentation. Similarly, for $\mathcal{L}_e$ and $\mathcal{L}_f$ at NC branch, for a given pixel i, we substitute X_i for $\Phi_i(I)$ and Y_i for $r_i(I, w_0, w_3)$ in equations (4) and (5). K is set as 5 within equation (4) when calculating $\mathcal{L}_e$, denoting the 4 types of nuclei that our model currently predicts and the background. Note, the value of K is chosen to reflect the number of nuclear types represented in the training set.

It must be noted that the NC branch loss $\mathcal{L}_e$ and $\mathcal{L}_f$ are only calculated when the classification labels are available. In other words, as mentioned in Section 3.1, the network performs only instance segmentation if there are no classification labels given.

3.2. Post Processing

Within each horizontal and vertical map, pixels between separate instances have a significant difference. This can be seen in Fig. 3 and is highlighted by the arrows. Therefore, calculating the gradient can inform where the nuclei should be separated because the output will give high values between neighbouring nuclei, where there is a significant difference in the pixel values. We define:

$$\mathcal{S}_m = max(H_x(p_x), H_y(p_y)) \tag{6}$$

where p_x and p_y refer to the the horizontal and vertical predictions at the output of the HoVer branch and H_x and H_y refer to the horizontal and vertical components of the Sobel operator. Specifically, H_x and H_y compute the horizontal and vertical derivative approximations and are shown by the gradient maps in Fig. 1. Therefore, $\mathcal{S}_m$ highlights areas where there is a significant difference in neighbouring pixels within the horizontal and vertical maps. Therefore, areas such as the ones shown by the arrows in Fig. 3 will result in high values within $\mathcal{S}_m$. We compute markers $M = \sigma(\tau(q,h) - \tau(S_m,k))$. Here, $\tau(a,b)$ is a threshold function that acts on a and sets values above b to 1 or 0 otherwise. Specifically, h and k were chosen such that they gave the optimal nuclear segmentation results. σ is a rectifier that sets all negative values to 0 and q is the probability map output of the NP branch. We obtain the energy landscape $E = [1 - \tau(S_m,k)] * \tau(q,h)$. Finally, M is used as the marker during marker-controlled watershed to determine how to split $\tau(q,h)$, given the energy landscape E. This sequence of events can be seen in Fig. 1.

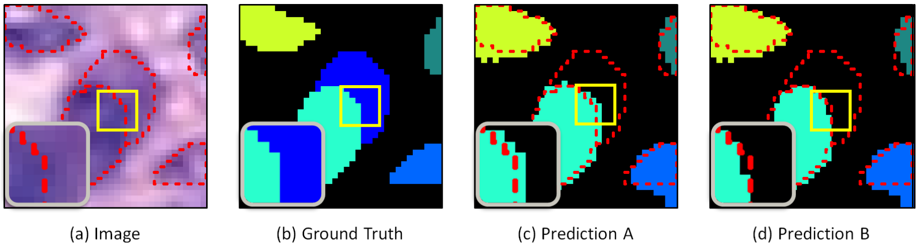
To perform simultaneous nuclear instance segmentation and classification, it is necessary to convert the per-pixel nuclear type prediction at the output of the NC branch to a prediction per nuclear instance. For each nuclear instance, we use *majority class* of the predictions made by the NC branch, i.e., the nuclear type of all pixels in an instance is assigned to be the class with the highest frequency count for that nuclear instance.

Please refer to Appendix A for a full analysis on the contribution of our proposed loss function, post-processing method and devoted classification branch.

4. Evaluation Metrics

4.1. Nuclear Instance Segmentation Evaluation

Assessment and comparison of different methods is usually given by an overall score that indicates which method is superior. However, to further investigate the method, it is preferable to break the problem into sub-tasks and measure the performance of the method on each sub-task. This enables an in depth analysis, thus facilitating a comprehensive understanding of the approach, which can help drive forward model development. For nuclear instance segmentation, the problem can be divided into the following three sub-tasks:


(a) Image

(b) Ground Truth

(c) Prediction A

(d) Prediction B

Figure 4: Examples highlighting the limitations of DICE2 and AJI with slightly different predictions. For better visualisation, ground truth contours (red dash line) for each instance have been overlaid on both the predictions and original images.

Table 1: Comparison between Prediction A and Prediction B from Fig.4 across various measurements.

	DICE2	**AJI**	**PQ**
Prediction A	0.6477	0.4790	0.6803
Prediction B	0.9007	0.6414	0.6863

- Separate the nuclei from the background
- Detect individual nuclear instances
- Segment each detected instance

In the current literature, two evaluation metrics have been mainly adopted to quantitatively measure the performance of nuclear instance segmentation: 1) Ensemble Dice (DICE2) (Vu et al., 2018), and 2) Aggregated Jaccard Index (AJI) (Kumar et al., 2017). Given the ground truth X and prediction Y, DICE2 computes and aggregates DICE per nucleus, where Dice coefficient (DICE) is defined as $2 \times (X \cap Y)/(|X|+|Y|)$ and AJI computes the ratio of an aggregated intersection cardinality and an aggregated union cardinality between X and Y.

These two evaluation metrics only provide an overall score for the instance segmentation quality and therefore provides no further insight into the sub-tasks at hand. In addition, these two metrics have a limitation, which we illustrate in Fig. 4. From the figure, although prediction A only differs from prediction B by a few pixels, the DICE2 and AJI scores for B are inferior. These scores are shown in Table 1. This problem arises due to over-penalisation of the overlapping regions. By overlaying the GT segment contours (red dashed line) upon the two predictions, we observe that, although the cyan-coloured instance within prediction A overlaps mostly with the cyan-coloured GT instance, it also slightly overlaps with the blue-coloured GT instance. As a result, according to the DICE2 algorithm, the predicted cyan instance will be penalised by pixels not only coming from the dominant overlapping cyan-coloured GT instance, but also from the blue-coloured GT instance. The AJI also suffers from the same phenomenon. However, because AJI only uses the prediction and GT instance pair with the highest intersection over union, over-penalisation is less likely compared to DICE2. Over-penalisation is likely to occur when the model completely fails to detect the neighbouring instance, such as in Fig. 4. Nonetheless, when evaluating methods across different datasets, specifically on samples containing lots of hard to recognise nuclei such as fibroblasts or nuclei with poor staining, the number of failed detections may increase and therefore may have a negative impact on the AJI measurement. Due to the limitations of DICE2 and AJI, it is clear that there is a need for an improved reliable quantitative measurement.

Panoptic Quality: We propose to use another metric for accurate quantification and interpretability to assess the performance of nuclear instance segmentation. Originally proposed by Kirillov et al. (2018), panoptic quality (PQ) for nuclear instance segmentation is defined as:

$$\mathcal{PQ} = \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Detection Quality(DQ)}} \times \underbrace{\frac{\sum_{(x,y)\in TP} IoU(x,y)}{|TP|}}_{\text{Segmentation Quality(SQ)}} \tag{7}$$

where x denotes a GT segment, y denotes a prediction segment and IoU denotes intersection over union. Each (x,y) pair is mathematically proven to be *unique* (Kirillov et al., 2018) over the entire set of prediction and GT segments if their IoU(x,y)>0.5. The unique matching splits all available segments into matched pairs (TP), unmatched GT segments (FN) and unmatched prediction segments

(FP). From this, PQ can be intuitively analysed as follows: the *detection quality* (DQ) is the F_1 Score that is widely used to evaluate instance detection, while *segmentation quality* (SQ) can be interpreted as how close each correctly detected instance is to their matched GT. DQ and SQ, in a way, also provide a direct insight into the second and third sub-tasks, defined above. We believe that PQ should set the standard for measuring the performance of nuclear instance segmentation methods.

Overall, to fully characterise and understand the performance of each method, we use the following three metrics: 1) DICE to measure the separation of all nuclei from the background; 2) Panoptic Quality as a unified score for comparison and 3) AJI for direct comparison with previous publications[3]. Panoptic quality is further broken down into DQ and SQ components for interpretability. Note, SQ is calculated only within true positive segments and should therefore be observed together with DQ. Throughout this study, these metrics are calculated for each image and the average of all images are reported as final values for each dataset.

4.2. Nuclear Classification Evaluation

Classification of the type of each nucleus is performed within the nuclear instances extracted from the instance segmentation or detection tasks. Therefore, the overall measurement for nuclear type classification should also encompass these two tasks. For all nuclear instances of a particular type t from both the ground truth and the prediction, the detection task d splits the GT and predicted instances into the following subsets: correctly detected instances (TP_d), misdetected GT instances (FN_d) and overdetected predicted instances (FP_d). Subsequently, the classification task c further breaks TP_d into correctly classified instances of type t (TP_c), correctly classified instances of types other than type t (TN_c), incorrectly classified instances of type t (FP_c) and incorrectly classified instances of types other than type t (FN_c). We then define the F_c

[3] Evaluation code available at: https://github.com/vqdang/hover_net/src/metrics

score of each type t for combined nuclear type classification and detection as follows:

$$F_c^t = \frac{2(TP_c + TN_c)}{2(TP_c + TN_c) + \alpha_0 FP_c + \alpha_1 FN_c + \alpha_2 FP_d + \alpha_3 FN_d} \tag{8}$$

where we use $\alpha_0 = \alpha_1 = 2$ and $\alpha_2 = \alpha_3 = 1$ to give more emphasis to nuclear type classification. Moreover, using the same weighting, if we further extend t to encompass all types of nuclei T ($t \in T$), the classification within TP_d is then divided into a correctly classified set A_c and an incorrectly classified set B_c. We can therefore disassemble F_c^t into:

$$\begin{aligned} F_c^T &= \frac{2A_c}{2(A_c + B_c) + FP_d + FN_d} = \frac{2(A_c + B_c)}{2(A_c + B_c) + FP_d + FN_d} \times \frac{A_c}{A_c + B_c} \\ &= F_d \times \text{Classification Accuracy within Correctly Detected Instances} \end{aligned} \tag{9}$$

where F_d is simply the standard detection quality like DQ while the other term is the accuracy of nuclear type classification within correctly detected instances. In the case where the GT is not exhaustively annotated for nuclear type classification, like in CRCHisto, an amount equal to the number of unlabelled GT instances in each set is subtracted from B_c and FN_c.

Finally, while IoU is utilised as the criteria in DQ for selecting the TP for detection in instance segmentation, detection methods can not calculate the IoU. Therefore, to facilitate comparison of both instance segmentation and detection methods for the nuclear type classification tasks, for F_c^t, we utilise the notion of distance to determine whether nuclei have been detected. To be precise, we define the region within a predefined radius from the annotated centre of the nucleus as the ground truth and if a prediction lies within this area, then it is considered to be a true positive. Here, we are consistent with Sirinukunwattana et al. (2016) and use a radius of 6 pixels at 20× or 12 pixels at 40×.

Table 2: Summary of the datasets used in our experiments. UHCW denotes University Hospitals Conventry and Warwickshire and TCGA denotes The Cancer Genome Atlas. *Seg* denotes segmentation masks and *Class* denotes classification labels.

	CoNSeP	**Kumar**	**CPM-15**	**CPM-17**	**TNBC**	**CRCHisto**
Total Number of Nuclei	24,319	21,623	2,905	7,570	4,056	29,756
Labelled Nuclei	24,319	0	0	0	0	22,444
Number of Images	41	30	15	32	50	100
Origin	UHCW	TCGA	TCGA	TCGA	Curie Institute	UHCW
Magnification	40×	40×	40× & 20×	40× & 20×	40×	20×
Size of Images	1000×1000	1000×1000	400×400 to 1000×600	500×500 to 600×600	512×512	500×500
Seg/Class	*Both*	*Seg*	*Seg*	*Seg*	*Seg*	*Class*
Number of Cancer Types	1	8	2	4	1	1

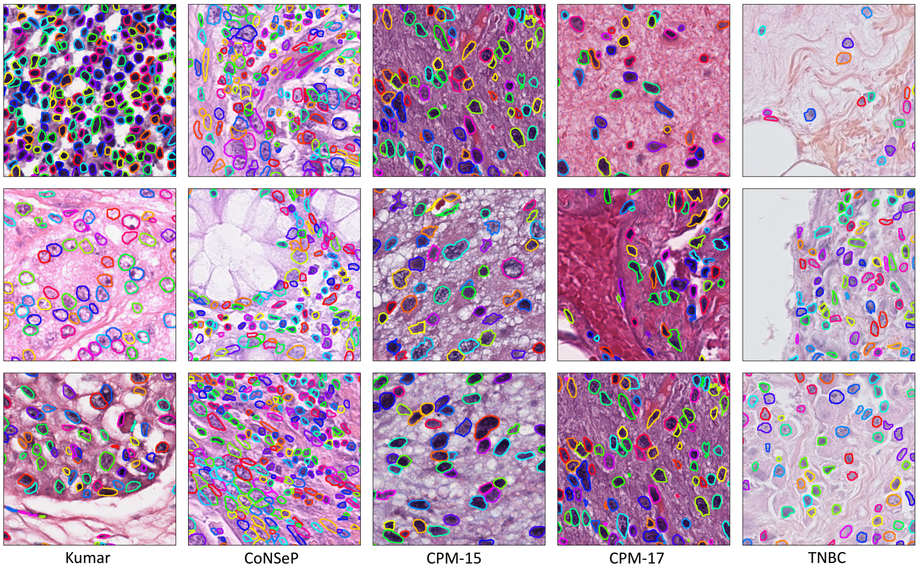


Figure 5: Sample cropped regions extracted from each of the five nuclear instance segmentation datasets used in our experiments. From left to right: Kumar (Kumar et al., 2017); CoNSeP; CPM-15; CPM-17 (Vu et al., 2018) and TNBC (Naylor et al., 2018). The different colours of nuclear contours highlight individual instances.

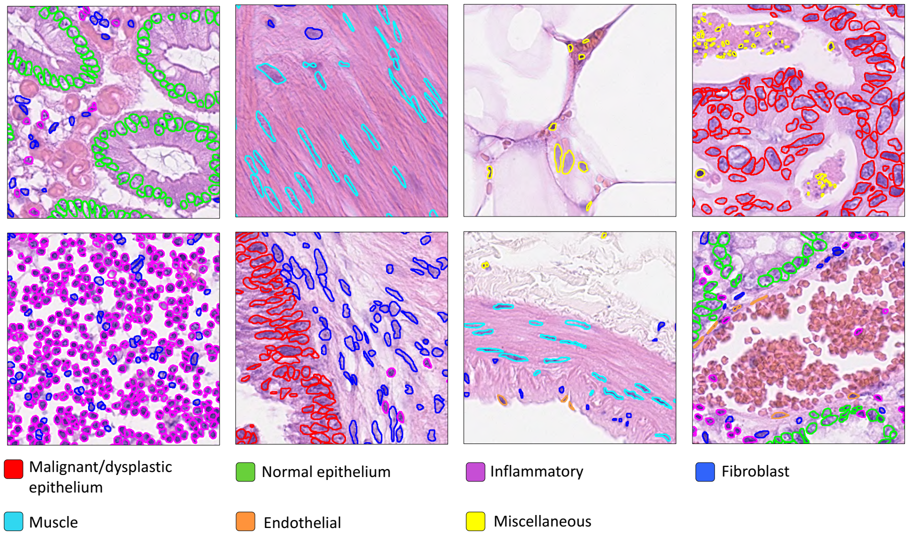


Figure 6: Sample cropped regions extracted from the CoNSeP datasets, where the colour of each nuclear boundary denotes the category.

5. Experimental Results

5.1. Datasets

As part of this work, we introduce a new dataset that we term as the colorectal nuclear segmentation and phenotypes (CoNSeP) dataset[4], consisting of 41 H&E stained image tiles, each of size 1,000×1,000 pixels at 40× objective magnification. Images were extracted from 16 colorectal adenocarcinoma (CRA) WSIs, each belonging to an individual patient, and scanned with an Omnyx VL120 scanner within the department of pathology at University Hospitals Coventry and Warwickshire, UK. We chose to focus on a *single* cancer type, so that we are able to display the true variation of tissue within colorectal adenocarcinoma WSIs, as opposed to other datasets that instead focus on using a small number of visual fields from various cancer types. Within this dataset, stroma, glandular, muscular, collagen, fat and tumour regions can be observed. Beside incorporating different tissue components, the 41 images were also cho-

[4] This dataset is available at https://warwick.ac.uk/fac/sci/dcs/research/tia/data/.

sen such that different nuclei *types* were present, including: normal epithelial; tumour epithelial; inflammatory; necrotic; muscle and fibroblast. Here, by *type* we are referring to the type of cell from which the nucleus originates from. Within the dataset, there are many significantly overlapping nuclei with indistinct boundaries and there exists various artifacts, such as ink. As a result of the diversity of the dataset, it is likely that a model trained on CoNSeP will perform well for unseen CRA cases. For each image tile, every nucleus was annotated by one of two expert pathologists (A.A, Y-W.T). After full annotation, each annotated sample was reviewed by *both* of the pathologists; therefore refining their own and each others' annotations. By the end of the annotation process, each pathologist had fully checked *every* sample and consensus had been reached. Annotating the data in this way ensured that minimal nuclei were missed in the annotation process. However, we can not avoid inevitable pixel-level differences between the annotation and the true nuclear boundary in challenging cases. In addition to delineating the nuclear boundaries, every nucleus was labelled as either: normal epithelial, malignant/dysplastic epithelial, fibroblast, muscle, inflammatory, endothelial or miscellaneous. Within the miscellaneous category, necrotic, mitotic and cells that couldn't be categorised were grouped. For our experiments, we grouped the normal and malignant/dysplastic epithelial nuclei into a single class and we grouped the fibroblast, muscle and endothelial nuclei into a class named spindle-shaped nuclei.

Overall, six independent datasets are utilised for this study. A full summary for each of them is provided in Table 2. Five of these datasets are used to evaluate the instance segmentation performance which we refer to as: CoNSeP; Kumar (Kumar et al., 2017); CPM-15; CPM-17 (Vu et al., 2018) and TNBC (Naylor et al., 2018). Example images from each of the five datasets can be seen in Fig. 7. Meanwhile, we utilise CoNSeP and a further dataset, named CRCHisto, to quantify the performance of the nuclear classification model. The CRCHisto dataset consists of the same nuclei types that are present in CoNSeP. It is also worth noting that the CRCHisto dataset is not exhaustively annotated for nuclear class labels.

5.2. Implementation and Training Details

We implemented our framework with the open source software library TensorFlow version 1.8.0 (Abadi et al., 2016) on a workstation equipped with two NVIDIA GeForce 1080 Ti GPUs. During training, data augmentation including flip, rotation, Gaussian blur and median blur was applied to all methods. All networks received an input patch with a size ranging from 252×252 to 270×270. This size difference is due to the use of valid convolutions in some architectures, such as HoVer-Net and U-Net. Regarding HoVer-Net, we initialised the model with pre-trained weights on the ImageNet dataset (Deng et al., 2009), trained only the decoders for the first 50 epochs, and then fine-tuned all layers for another 50 epochs. We train stage one for around 120 minutes and stage two for around 260 minutes. Therefore, the overall training time is around 380 minutes. Stage two takes longer to train because unfreezing the encoder utilises more memory and therefore a smaller batch size needs to be used. Specifically, we used a batch size of 8 and 4 on each GPU for stage one and two respectively. We used Adam optimisation with an initial learning rate of 10^{-4} and then reduced it to a rate of 10^{-5} after 25 epochs. This strategy was repeated for fine-tuning. On the whole, training of the network is stable, where the usage of fully independent decoders helps the network to converge each time. The network was trained with an RGB input, normalised between 0 and 1.

5.3. Comparative Analysis of Segmentation Methods

Experimental Setting: We evaluated our approach by employing a full independent comparison across the three largest known exhaustively labelled nuclear segmentation datasets: Kumar; CoNSeP and CPM-17 and utilised the metrics as described in Section 4.1. For this experiment, because we do not have the classification labels for all datasets, we perform instance segmentation without classification. This enables us to fully leverage all data and allows us to rigorously evaluate the segmentation capability of our model. In the same way as Kumar et al. (2017), we split the Kumar dataset into two different sub-datasets: (i) Kumar-Train, a training set with 16 image tiles (4 breast, 4 liver,

Table 3: Comparative experiments on the Kumar (Kumar et al., 2017), CoNSeP and CPM-17 (Vu et al., 2018) datasets. WS denotes watershed-based post processing.

	Kumar					**CoNSeP**					**CPM-17**				
Methods	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**
Cell Profiler (Carpenter et al., 2006)	0.623	0.366	0.423	0.704	0.300	0.434	0.202	0.249	0.705	0.179	0.570	0.338	0.368	0.702	0.261
QuPath (Bankhead et al., 2017)	0.698	0.432	0.511	0.679	0.351	0.588	0.249	0.216	0.641	0.151	0.693	0.398	0.320	0.717	0.230
FCN8 (Long et al., 2015)	0.797	0.281	0.434	0.714	0.312	0.756	0.123	0.239	0.682	0.163	0.840	0.397	0.575	0.750	0.435
FCN8 + WS (Long et al., 2015)	0.797	0.429	0.590	0.719	0.425	0.758	0.226	0.320	0.676	0.217	0.840	0.397	0.575	0.750	0.435
SegNet (Badrinarayanan et al., 2017)	0.811	0.377	0.545	0.742	0.407	0.796	0.194	0.371	0.727	0.270	0.857	0.491	0.679	0.778	0.531
SegNet + WS (Badrinarayanan et al., 2017)	0.811	0.508	0.677	0.744	0.506	0.793	0.330	0.464	0.721	0.335	0.856	0.594	0.779	0.784	0.614
U-Net (Ronneberger et al., 2015)	0.758	0.556	0.691	0.690	0.478	0.724	0.482	0.488	0.671	0.328	0.813	0.643	0.778	0.734	0.578
Mask-RCNN (He et al., 2017)	0.760	0.546	0.704	0.720	0.509	0.740	0.474	0.619	0.740	0.460	0.850	0.684	0.848	0.792	0.674
DCAN (Chen et al., 2016)	0.792	0.525	0.677	0.725	0.492	0.733	0.289	0.383	0.667	0.256	0.828	0.561	0.732	0.740	0.545
Micro-Net (Raza et al., 2018)	0.797	0.560	0.692	0.747	0.519	0.794	0.527	0.600	0.745	0.449	0.857	0.668	0.836	0.788	0.661
DIST (Naylor et al., 2018)	0.789	0.559	0.601	0.732	0.443	0.804	0.502	0.544	0.728	0.398	0.826	0.616	0.663	0.754	0.504
CNN3 (Kumar et al., 2017)	0.762	0.508	-	-	-	-	-	-	-	-	-	-	-	-	-
CIA-Net (Zhou et al., 2019)	0.818	**0.620**	0.754	0.762	0.577	-	-	-	-	-	-	-	-	-	-
DRAN (Vu et al., 2018)	-	-	-	-	-	-	-	-	-	-	0.862	0.683	0.811	0.804	0.657
HoVer-Net	**0.826**	0.618	**0.770**	**0.773**	**0.597**	**0.853**	**0.571**	**0.702**	**0.778**	**0.547**	**0.869**	**0.705**	**0.854**	**0.814**	**0.697**

4 kidney and 4 prostate) and (ii) Kumar-Test, a test set with 14 image tiles (2 breast, 2 liver, 2 kidney and 2 prostate, 2 bladder, 2 colon, 2 stomach). Note, we utilise the exact same image split used by other recent approaches (Kumar et al., 2017; Naylor et al., 2018; Zhou et al., 2019), but we do not separate the test set into two subsets. We do this to ensure that the test set is large enough, ensuring a reliable evaluation. For CoNSeP, we devise a suitable train and test set that contains 26 and 14 images respectively. The images within the test set were selected to ensure the true diversity of nuclei types within colorectal tissue are represented. For CPM-17, we utilise the same split that had been employed for the challenge, with 32 images in both the training and test datasets.

We compared our proposed model to recent segmentation approaches used in computer vision (Long et al., 2015; Badrinarayanan et al., 2017; He et al., 2017), medical imaging (Ronneberger et al., 2015) and also to methods specifically tuned for the task of nuclear segmentation (Chen et al., 2016; Raza et al., 2018; Naylor et al., 2018; Zhou et al., 2019; Vu et al., 2018). We also compared the performance of our model to two open source software applications: Cell Profiler (Carpenter et al., 2006) and QuPath (Bankhead et al., 2017). Cell Profiler is a software for cell-based analysis, with several suggested pipelines for

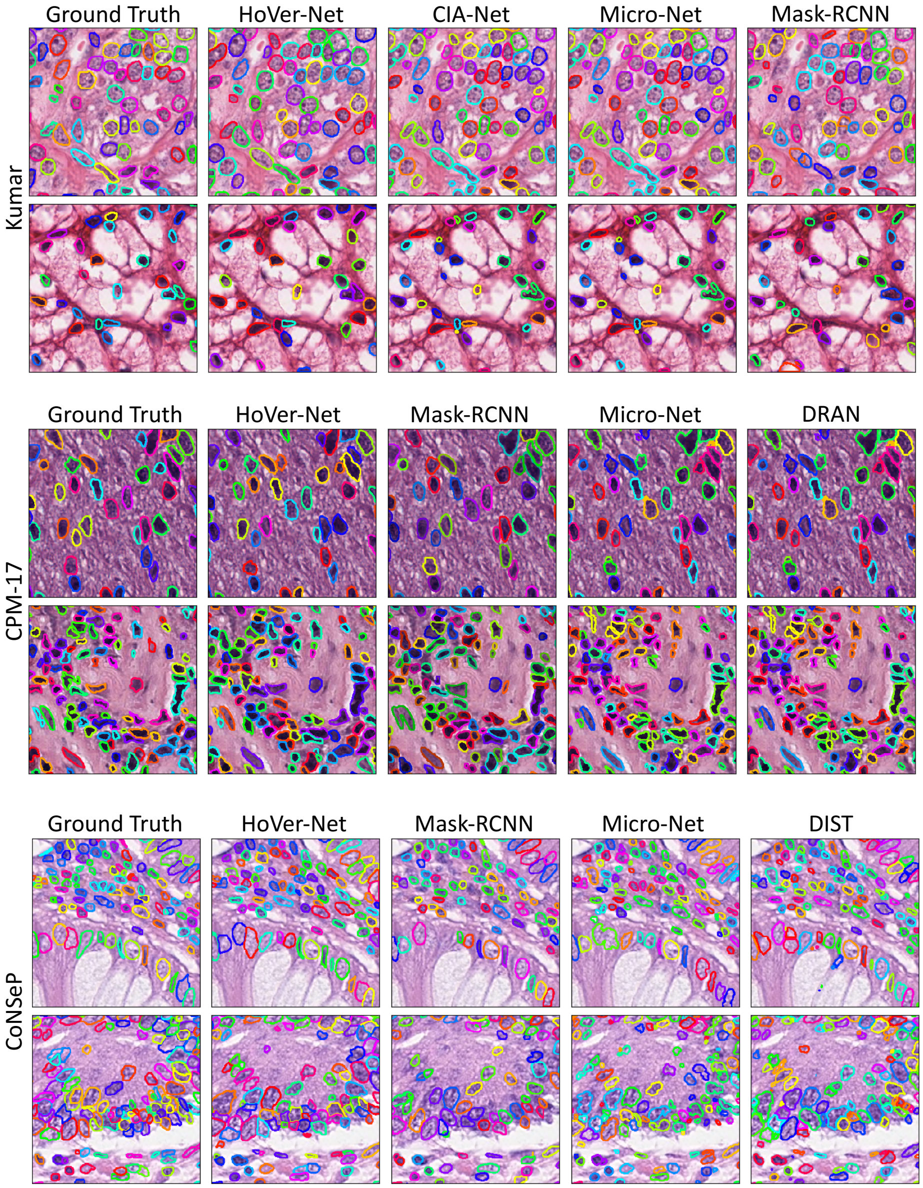
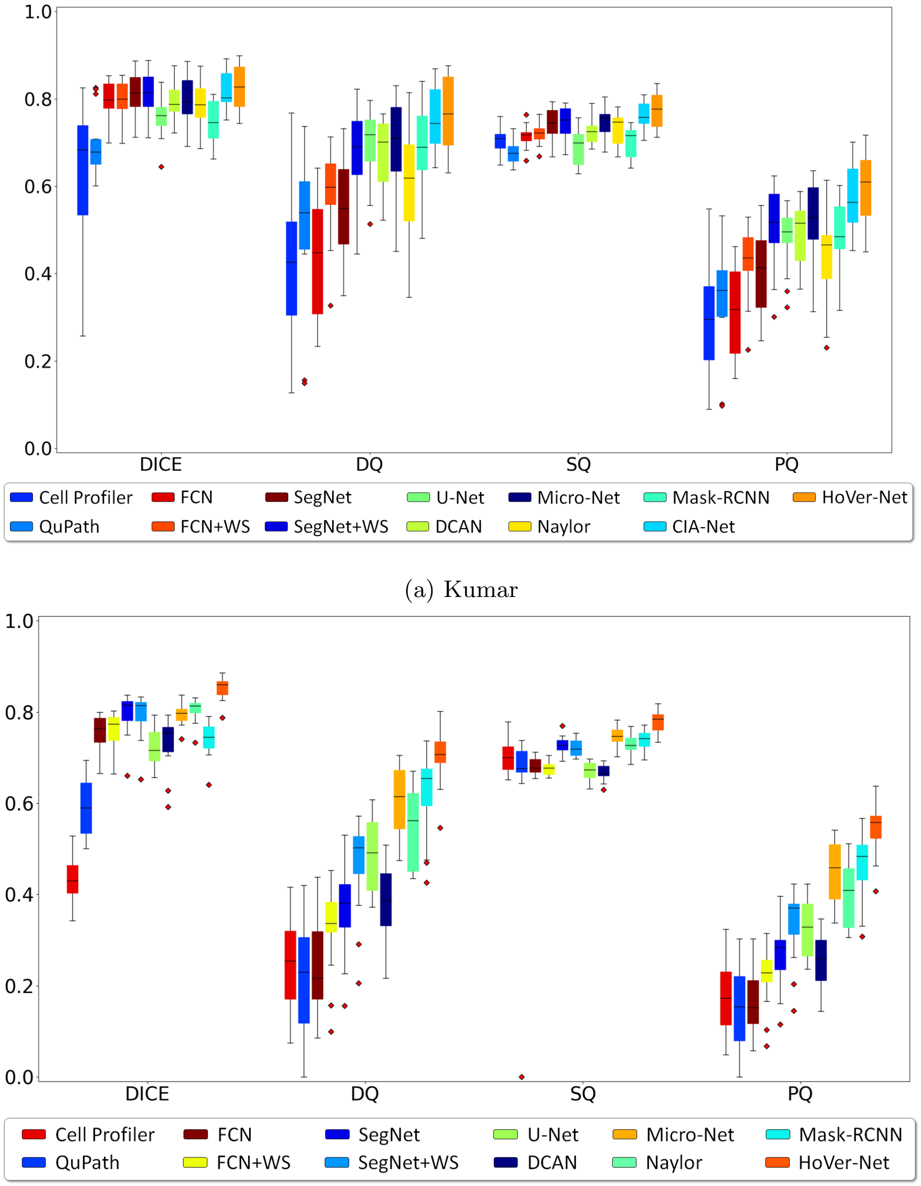


Figure 7: Example visual results on the CPM-17, Kumar and CoNSeP datasets. For each dataset, we display the 4 models that achieve the highest PQ score from left to right. The different colours of the nuclear boundaries denote separate instances.

500 computational pathology. The pipeline that we adopted applies a threshold to the greyscale image and then uses a series of post processing operations. QuPath

is an open source software for digital pathology and whole slide image analysis. To achieve nuclear segmentation, we used the default parameters within the application. FCN, SegNet, U-Net, DCAN, Mask-RCNN and DIST have been implemented by the authors of the paper (S.G, Q.D.V). For Mask-RCNN, we slightly modified the original implementation by using smaller anchor boxes. The default configuration is fine-tuned for natural images and therefore, this modification was necessary to perform a successful nuclear segmentation. DIST was implemented with the assistance of the first author of the corresponding approach in order to ensure reliability during evaluation. This also enabled us to utilise DIST for further comparison in our experiments. For Micro-Net, we used the same implementation that was described by Raza et al. (2018) and was implemented by the first author of the corresponding paper (S.E.A.R). For CNN3 and CIA-Net, we report the results on the Kumar dataset that are given in their respective original papers. The authors of CIA-Net and DRAN provided their segmentation output, which meant that we were able to obtain all metrics on the datasets that the models were applied to. Therefore, we report results of CIA-Net on the Kumar dataset and results of DRAN on the CPM-17 dataset. Note, for all self-implemented approaches we are consistent with our pre-processing strategy. However, DRAN, CNN3 and CIA-Net results are directly taken from their respective papers and therefore we can't guarantee the same pre-processing steps. CNN3 and CIA-Net also use stain normalisation, whereas other methods described in this paper do not.

Comparative Results: Table 3 and the box plots in Fig. 8a and 8b show detailed results of this experiment. Within the box plots, we choose not to show AJI, due to its limitations as discussed in Section 4.1. A large variation in performance between methods within each dataset is observed. This variation is particularly evident in the Kumar and CoNSeP datasets, where there exists a large number of overlapping nuclei. Both Cell Profiler (Carpenter et al., 2006) and QuPath (Bankhead et al., 2017) achieve sub-optimal performance for all datasets. In particular, both software applications consistently achieve a low DICE score, suggesting that their inability to distinguish nuclear pixels from



(a) Kumar

(b) CoNSeP

Figure 8: Box plots highlighting the performance of competing methods on the Kumar and CoNSeP datasets.

the background is a major limiting factor. FCN-based approaches improve the capability of models to detect nuclear pixels, yet often fail due to their inability to separate clustered instances. For example, despite a higher DICE score than Cell Profiler and QuPath, networks built only for semantic segmentation like FCN8 and SegNet suffer from low PQ values. Therefore, methods that incorporate strong instance-aware techniques are favourable. Within CPM-17, there are less overlapping nuclei which explains why methods that are not instance-aware are still able to achieve a satisfactory performance. We observe that the weighted cross entropy loss that is used in both U-Net and Micro-Net can help to separate joined nuclei, but its success also depends on the capacity of the network. This is reflected by the increased performance of Micro-Net over U-Net.

DCAN is able to better distinguish between separate instances than FCN8, which uses a very similar encoder based on the VGG16 network. Therefore, incorporating additional information at the output of the network can improve the segmentation performance. This is also exemplified by the fairly strong performances of CNN3, DIST, DRAN and CIA-Net. In a different way, Mask-RCNN is able to successfully separate clustered nuclei by utilising a region proposal based approach. However, Mask-RCNN is less effective than other methods at detecting nuclear pixels, which is reflected by a lower DICE score.

Due to the reasoning given in Section 4, we place a larger emphasis on PQ to determine the success of different models. In particular, we consistently obtain an improved performance over DIST, which justifies the use of our proposed horizontal and vertical maps as a regression target. We also report a better performance than the winners of the Computational Precision Medicine and MoNuSeg challenges (Vu et al., 2018; Zhou et al., 2019), that utlised the CPM-17 and Kumar datasets respectively. Therefore, HoVer-Net achieves state-of-the art performance for nuclear instance segmentation compared to all competing methods on multiple datasets that consist of a variety of different tissue types. Our approach also outperforms methods that were fine-tuned for the task of nuclear segmentation.

5.4. Generalisation Study

Experimental Setting: The goal of any automated method is to perform well on unseen data, with high accuracy. Therefore, we conducted a large scale study to assess how all methods generalise to new H&E stained images. To analyse the generalisation capability, we assessed the ability to segment nuclei from: i) new organs (variation in nuclei shapes) and ii) different centres (variation in staining).

The five instance segmentation datasets used within our experiments can be grouped into three groups according to their origin: TCGA (Kumar, CPM-15, CPM-17), TNBC and CoNSeP. We used Kumar as the training and validation set, due to its size and diversity, whilst the combined CPM (CPM-15 and CPM-17), TNBC and CoNSeP datsets were used as three independent test sets. We split the test sets in this way in accordance with their origin. Note, for this experiment we use both the training and test sets of CPM-17 and CoNSeP to form the independent test sets. Kumar was split into three subsets, as explained in Section 5.1, and Kumar-Train was used to train all models, i.e. trained with samples originating from the following organs: breast; prostate; kidney and liver. Despite all samples being extracted from TCGA, CPM samples come from the brain, head & neck and lungs regions. Therefore, testing with CPM reflects the ability for the model to generalise to new organs, as mentioned above by the first generalisation criterion. TNBC contains samples from an already seen organ (breast), but the data is extracted from an independent source with different specimen preservation and staining practice. Therefore, this reflects the second generalisation criterion. CoNSeP contains samples taken from colorectal tissue, which is not represented in Kumar-Train, and is also extracted from a source independent to TCGA. Therefore, this reflects *both* the first and second generalisation criteria. Also, as mentioned in Section 5.1, CoNSeP contains challenging samples, where there exists various artifacts and there is variation in the quality of slide preparation. Therefore, the performance on this dataset also reflects the ability of a model to generalise to difficult samples.

Comparative Results: The results are reported in Table 4, where we only

Table 4: Comparative results, highlighting the generalisation capability of different models. All models are initially trained on Kumar and then the Combined CPM (Vu et al., 2018), TNBC (Naylor et al., 2018) and CoNSeP datasets are processed.

	Combined CPM					TNBC					All CoNSeP				
Methods	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**
FCN8 + WS (Long et al., 2015)	0.762	0.531	0.669	0.722	0.487	0.726	0.506	0.662	0.723	0.480	0.609	0.247	0.345	0.688	0.240
SegNet + WS (Badrinarayanan et al., 2017)	0.791	0.583	0.738	0.755	0.561	0.758	0.559	0.734	0.750	0.554	0.681	0.315	0.449	0.733	0.332
U-Net (Ronneberger et al., 2015)	0.720	0.541	0.652	0.672	0.446	0.681	0.514	0.635	0.676	0.442	0.585	0.363	0.442	0.670	0.297
Mask-RCNN (He et al., 2017)	0.764	0.575	0.760	0.719	0.549	0.705	0.529	0.726	0.742	0.543	0.606	0.348	0.492	0.720	0.357
DCAN (Chen et al., 2016)	0.770	0.582	0.716	0.730	0.528	0.725	0.537	0.683	0.720	0.495	0.609	0.306	0.403	0.685	0.278
Micro-Net (Raza et al., 2018)	0.792	0.615	0.716	0.751	0.542	0.701	0.531	0.656	0.753	0.497	0.644	0.394	0.489	0.722	0.356
DIST (Naylor et al., 2018)	0.775	0.563	0.593	0.720	0.432	0.719	0.523	0.549	0.714	0.404	0.621	0.369	0.379	0.701	0.268
HoVer-Net	**0.801**	**0.626**	**0.774**	**0.778**	**0.606**	**0.749**	**0.590**	**0.743**	**0.759**	**0.578**	**0.664**	**0.404**	**0.529**	**0.764**	**0.408**

display the results of methods that employ an instance-based technique. We observe that our proposed model is able to successfully generalise to unseen data in all three cases. However, some methods prove to perform poorly with unseen data, where in particular, U-Net and DIST perform worse than other competing methods on all three datasets. Both SegNet with watershed and Mask-RCNN achieve a competitive performance across all three generalisation tests. However, similar to the results reported in Table 3, Mask-RCNN is not able to distinguish nuclear pixels from the background as well as other competing methods, which has an adverse effect on the overall segmentation performance shown by PQ. On the other hand, SegNet proves to successfully detect nuclear pixels, reporting a greater DICE score than HoVer-Net on both the TNBC and CoNSeP datasets. However, the overall segmentation result for HoVer-Net is superior because it is better able to separate nuclear instances by incorporating the horizontal and vertical maps at the output of the network.

5.5. Comparative Analysis of Classification Methods

Experimental Setting: We converted the top four performing nuclear instance segmentation algorithms, based on their panoptic quality on the CoNSeP dataset, such that they were able to perform simultaneous instance segmentation and classification. As mentioned in Section 5.1, the nuclear categories that we use in our experiments are: miscellaneous, inflammatory, epithelial and spindle-shaped. Specifically, we compared HoVer-Net with Micro-Net, Mask-RCNN and DIST. For Micro-Net, we used an output depth of 5 rather than 2, where each channel gave the probability of a pixel being either background, miscellaneous, inflammatory, epithelial or spindle-shaped. For Mask-RCNN, there is a devoted classification branch that predicts the class of each instance and therefore is well suited to a multi-class setting. DIST performs regression at the output of the network and therefore converting the model such that it is able to classify nuclei into multiple categories is non-trivial. Instead, we add an extra 1×1 convolution at the output of the network that performs nuclear classification. As well as comparing to the aforementioned methods, we compared our approach to a spatially constrained CNN (SC-CNN), that achieves detection and classification. Note, because SC-CNN does not produce a segmentation mask, we do not report the PQ for this method.

Comparative Results: We trained our models on the training set of the CoNSeP dataset and then we evaluated the model on both the test set of CoNSeP and also the entire CRCHisto dataset. Table 5 displays the results of the multi-class models on the CoNSeP and the CRCHisto datasets respectively, where the given metrics are described in Section 4.2. For CoNSeP, along with the classification metrics, we provide PQ as an indication of the quality of instance segmentation. However, in CRCHisto, only the nuclear centroids are given and therefore, we exclude PQ from the CRCHisto evaluation because it can't be calculated without the instance segmentation masks. We observe that HoVer-Net achieves a good quality simultaneous instance segmentation and classification, compared to competing methods. It must be noted, that we should expect a lower F_1 score for the miscellaneous class because there are significantly less nuclei represented. Also, there is a high diversity of nuclei types that have been grouped within this class, belonging to: mitotic; necrotic and cells that are uncategorisable. Despite this, HoVer-Net is able to achieve a satisfactory performance on this class, where other methods fail. Furthermore, compared to other methods, our approach achieves the best F_1 score for epithelial, inflammatory and spindle classes. Therefore, due to HoVer-Net obtaining a strong performance for both nuclear segmentation and classification, we suggest that our model may be used for sophisticated subsequent cell-level downstream analysis in computational pathology.

Table 5: Comparative results for nuclear classification on the CoNSeP and CRCHisto datasets. F_d denotes the F_1 score for nuclear detection, whereas F_c^e, F_c^i, F_c^s and F_c^m denote the F_1 classification score for the epithelial, inflammatory, spindle-shaped and miscellaneous classes respectively.

	CoNSeP						**CRCHisto**				
Methods	**PQ**	$\mathbf{F}_d$	$\mathbf{F}_c^e$	$\mathbf{F}_c^i$	$\mathbf{F}_c^s$	$\mathbf{F}_c^m$	$\mathbf{F}_d$	$\mathbf{F}_c^e$	$\mathbf{F}_c^i$	$\mathbf{F}_c^s$	$\mathbf{F}_c^m$
SC-CNN (Sirinukunwattana et al., 2016)	-	0.608	0.306	0.193	0.175	0.000	0.664	0.246	0.111	0.126	0.000
DIST (Naylor et al., 2018)	0.372	0.712	0.617	0.534	0.505	0.000	0.616	0.464	0.514	0.275	0.000
Micro-Net (Raza et al., 2018)	0.430	0.743	0.615	0.592	0.532	0.117	0.638	0.422	0.518	0.249	0.059
Mask-RCNN (He et al., 2017)	0.450	0.692	0.595	0.590	0.520	0.098	0.639	**0.503**	0.537	0.294	0.077
HoVer-Net	**0.516**	**0.748**	**0.635**	**0.631**	**0.566**	**0.426**	**0.688**	0.486	**0.573**	**0.302**	**0.178**

6. Discussion and Conclusions

Analysis of nuclei in large-scale histopathology images is an important step towards automated downstream analysis for diagnosis and prognosis of cancer. Nuclear features have been often used to assess the degree of malignancy (Gurcan et al., 2009). However, visual analysis of nuclei is a very time consuming task because there are often tens of thousands of nuclei within a given whole-slide image (WSI). Performing simultaneous nuclear instance segmentation and classification enables subsequent exploration of the role that nuclear features play in predicting clinical outcome. For example, Lu et al. (2018) utilised nuclear features from histology TMA cores to predict survival in early-stage estrogen receptor-positive breast cancer. Restricting the analysis to some specific nuclear types only may be advantageous for accurate analysis in computational pathology.

In this paper, we have proposed HoVer-Net for simultaneous segmentation and classification of nuclei within multi-tissue histology images that not only detects nuclei with high accuracy, but also effectively separates clustered nuclei. Our approach has three up-sampling branches: 1) the nuclear pixel branch that separates nuclear pixels from the background; 2) the HoVer branch that regresses the horizontal and vertical distances of nuclear pixels to their centres of mass and 3) the nuclear classification branch that determines the type of each nucleus. We have shown that the proposed approach achieves the state-of-the-art instance segmentation performance compared to a large number of recently published deep learning models across multiple datasets, including tissues that have been prepared and stained under different conditions. This makes the proposed approach likely to translate well to a practical setting due its strong generalisation capacity, which can therefore be effectively used as a prerequisite step before nuclear-based feature extraction. We have shown that utilising the horizontal and vertical distances of nuclear pixels to their centres of mass provides powerful instance-rich information, leading to state-of-the-art performance in histological nuclear segmentation. When the classification labels are available, we show that our model is able to successfully segment and classify nuclei with high accuracy.

Region proposal (RP) methods, such as Mask-RCNN, show great potential in dealing with overlapping instances because there is no notion of *separating* instances; instead nuclei are segmented independently. However, a major limitation of the RP methods is the difficulty in merging instance predictions between neigbouring tiles during processing. For example, if a sub-segment of a nucleus at the boundary is assigned a label, one must ensure that the remainder of the nucleus in the neighbouring tile is also assigned the same label. To overcome this difficulty, for Mask-RCNN, we utilised an overlapping tile mechanism such that we only considered non-boundary nuclei.

Regarding the processing time, the average time to process a 1,000×1,000 image tile over 10 runs using Mask-RCNN for segmentation and classification was 106.98 seconds. Meanwhile, HoVer-Net only took an average of 11.04 seconds to complete the same operation; approximately 9.7× faster. On the other hand, the average processing time for DIST and Micro-Net was 0.600 and 0.832

seconds respectively. Mask-RCNN inherently stores a single instance per channel, which leads to very large arrays in memory when there are many nuclei in a single image patch, which also contributes to the much longer processing time as seen above. Overall, FCN methods seem to better translate to WSI processing compared to Mask-RCNN or RPN methods in general. It must be stressed that the timing is not exact and is dependent on hardware specifications and software implementation. With optimised code and sophisticated hardware, we expect these timings to be considerably different. Additionally, the inference time is also dependent on the size of the output. In particular, with a smaller output size, a smaller stride is also required during processing. For instance, if we used padded convolution in the up-sampling branches of HoVer-Net, then we observe 5.6× speed up and the average processing time is 1.97 seconds per 1000×1000 image tile. For fair comparison, all models were processed on a single GPU with 12GB RAM and we fixed the batch size to a size of one. Future work will explore the trade-off between the efficiency of HoVer-Net and its potential to accurately perform instance segmentation and classification.

A major bottleneck for the development of successful nuclear segmentation algorithms is the limitation of data; particularly with additional associated class labels. In this work, we introduce the colorectal adenocarcinoma nuclear segmentation and phenotypes (CoNSeP) dataset, containing over 24K labelled nuclei from challenging samples to reflect the true difficulty of segmenting nuclei in whole-slide images. Due to the abundance of nuclei with an associated nuclear category, CoNSeP aims to help accelerate the development of further simultaneous nuclear instance segmentation and classification models to further increase the sophistication of cell-level analysis within computational pathology.

We analysed the common measurements used to assess the true performance of nuclear segmentation models and discussed their limitations. Due to the fact that these measurements did not always reflect the instance segmentation performance, we proposed a set of reliable and informative statistical measures. We encourage researchers to utilise the proposed measures to not only maximise the interpretability of their results, but also to perform a fair comparison with other methods.

Finally, methods have surfaced recently that explore the relationship of various nuclear types within histology images (Javed et al., 2018; Sirinukunwattana et al., 2018), yet these methods are limited to spatial analysis because the segmentation masks are not available. Utilising our model for nuclear segmentation and classification enables the exploration of the spatial relationship between various nuclear types combined with nuclear morphological features and therefore may provide additional diagnostic and prognostic value. Currently, our model is trained on a single tissue type, yet due to the strong performance of our instance segmentation model across multiple tissues, we are confident that our model will perform well if we were to incorporate additional tissue types. We observe a low F_1 classification score for the miscellaneous category in the classification model because there are significantly less samples within this category and there exists high intra-class variability. Future work will involve obtaining more samples within this category, including necrotic and mitotic nuclei, to improve the class balance of the data.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1C1B2012433) and by the Ministry of Science and ICT (MSIT) (No. 2018K1A3A1A74065728). We also acknowledge the financial support from EPSRC and MRC, provided as part of the Mathematics for Real-World Systems CDT. We thank Peter Naylor for his assistance in the implementation of the DIST network.

Appendix A. Ablation Studies

To gain a full understanding of the contribution of our method, we investigated several of its components. Specifically, we performed the following ablation experiments: (i) contribution of the proposed loss strategy; (ii) Sobel-based post processing technique compared to other strategies and (iii) contribution of the dedicated classification branch. Here, we utilised the Kumar and CoNSeP datasets for (i) and (ii) due to the large number of nuclei present, whereas for (iii) we use CoNSeP and CRCHisto because we do not have the classification labels for Kumar.

Loss Terms: We conducted an experiment to understand the contribution of our proposed loss strategy. First, we used mean squared error (MSE) of the horizontal and vertical distances L_a as the loss function of the HoVer branch and binary cross entropy (BCE) loss L_c as the loss function for the NP branch. We refer to this combination as the *standard* strategy because MSE and BCE are the two most commonly used loss functions for regression and binary classification tasks respectively. Next, we introduced the MSE of the horizontal and vertical gradients L_b to the HoVer branch and the dice loss L_d to the NP branch. The intuition behind our novel L_b is that it enforces the correct structure of the horizontal and vertical map predictions and therefore helps to correctly separate neighbouring instances. The dice loss was introduced because it can help the network to better distinguish between background and nuclear pixels and is particularly useful when there is a class-imbalance. We present the results in Table A1, where we observe an increase in all performance measures for our proposed multi-term loss strategy. Therefore, the additional loss terms boost the network's ability to differentiate between nuclear and background pixels (DICE) and separate individual nuclei (DQ and PQ). In particular, there is a significant boost in the SQ for both Kumar and CoNSeP, which suggests that our proposed loss function L_b is necessary to precisely determine where nuclei should be split.

Post Processing: Usually, markers obtained from applying a threshold to an energy landscape (such as the distance map) is enough to provide a competitive input for watershed, as seen by DIST in Table 3. Although HoVer-Net is not directly built upon an energy landscape, we devised a Sobel-based method to derive both the energy landscape and the markers. To compare with other methods, we implemented two further techniques for obtaining the energy landscape and the markers. We then exhaustively compared all energy landscape and marker combinations to assess which post processing strategy is the best. We start by linking HoVer to the distance map by calculating the square sum $\chi^2 + \varphi^2$, which can be seen as the distance from a pixel to its nearest nuclear centroid. In other words, this is a pseudo distance map. Additionally, χ and φ values can be interpreted as Cartesian coordinates with each nuclear centroid as the origin. By thresholding the values between a certain range, we can obtain the markers. The results of all combinations are shown in Table A2. Note, our gradient-based post processing technique is specifically designed for the HoVer branch output.

Classification Branch: In order to assess the importance of a devoted branch for concurrent nuclear segmentation and classification, we compared the proposed three branch setup of HoVer-Net to a two branch setup. Here, the two branch setup extends the NP branch to a multi-class setting, by predicting each nuclear type at the output. Then, to obtain the binary mask, the positive channels are combined together after nuclear type prediction. Utilising three branches decouples the tasks of nuclear classification and nuclear detection, where a separate branch is devoted to each task. For this ablation study, we train on the CoNSeP training set and then process both the CoNSeP test set and the entire CRCHisto dataset.

We report results in Table A3, where we observe that utilising a separate branch devoted to the task of nuclear classification leads to an improved overall performance of simultaneous nuclear instance segmentation and classification in both the CoNSeP and CRCHisto datasets. We can see that if the classification takes place at the output of NP branch, then the network's ability to determine the nuclear type is compromised. This is because the task of nuclear classification is challenging and therefore the network benefits from the introduction of a branch dedicated to the task of classification.

Table A1: Ablation study highlighting the contribution of the proposed loss strategy.

	Kumar					CoNSeP				
Strategy	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**
Standard Loss	0.823	0.750	0.771	0.581	0.608	0.846	0.685	0.774	0.532	0.557
Proposed Loss	**0.826**	**0.770**	**0.773**	**0.597**	**0.618**	**0.853**	**0.702**	**0.778**	**0.547**	**0.571**

Table A2: Ablation study for post processing techniques: Sobel-based versus thresholding to get markers and Sobel-based versus naive conversion to get energy landscape

		Kumar					CoNSeP				
Energy	**Markers**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**	**DICE**	**AJI**	**DQ**	**SQ**	**PQ**
$\chi^2 + \varphi^2$	Threshold	0.825	0.597	0.705	0.764	0.541	0.850	0.543	0.602	0.761	0.459
$\chi^2 + \varphi^2$	Sobel	0.826	0.613	0.766	0.768	0.591	0.853	0.561	0.694	0.770	0.535
Sobel	Threshold	0.825	0.614	0.715	0.772	0.554	0.850	0.566	0.617	0.775	0.479
Sobel	Sobel	**0.826**	**0.618**	**0.770**	**0.773**	**0.597**	**0.853**	**0.571**	**0.702**	**0.778**	**0.547**

Table A3: Ablation study showing the contribution of the classification branch in HoVer-Net on the CoNSeP dataset. F_d denotes the F_1 score for nuclear detection, whereas F_c^e, F_c^i, F_c^s and F_c^m denote the F_1 classification score for the epithelial, inflammatory, spindle-shaped and miscellaneous classes respectively.

	CoNSeP						CRCHisto				
Branches	**PQ**	$\mathbf{F}_d$	$\mathbf{F}_c^e$	$\mathbf{F}_c^i$	$\mathbf{F}_c^s$	$\mathbf{F}_c^m$	$\mathbf{F}_d$	$\mathbf{F}_c^e$	$\mathbf{F}_c^i$	$\mathbf{F}_c^s$	$\mathbf{F}_c^m$
NP & HoVer	0.499	0.736	**0.636**	0.545	0.528	0.333	0.666	0.458	0.523	0.271	0.132
NP & HoVer & NC	**0.516**	**0.748**	0.635	**0.631**	**0.566**	**0.426**	**0.688**	**0.486**	**0.573**	**0.302**	**0.178**

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning., in: OSDI, pp. 265–283.

Ali, S., Madabhushi, A., 2012. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. IEEE transactions on medical imaging 31, 1448–1460.

Alsubaie, N., Sirinukunwattana, K., Raza, S.E.A., Snead, D., Rajpoot, N., 2018. A bottom-up approach for tumour differentiation in whole slide images of lung adenocarcinoma, in: Medical Imaging 2018: Digital Pathology, International Society for Optics and Photonics. p. 105810E.

Arnab, A., Miksik, O., Torr, P.H.S., 2017. On the robustness of semantic segmentation models to adversarial attacks. CoRR abs/1711.09856. URL: http://arxiv.org/abs/1711.09856, arXiv:1711.09856.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 2481–2495.

Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al., 2017. Qupath: Open source software for digital pathology image analysis. Scientific reports 7, 16878.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al., 2006. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology 7, R100.

Chen, H., Qi, X., Yu, L., Heng, P.A., 2016. Dcan: deep contour-aware networks for accurate gland segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2487–2496.

Cheng, J., Rajapakse, J.C., et al., 2009. Segmentation of clustered nuclei with shape markers and marking function. IEEE Transactions on Biomedical Engineering 56, 741–748.

Corredor, G., Wang, X., Zhou, Y., Lu, C., Fu, P., Syrigos, K., Rimm, D.L., Yang, M., Romero, E., Schalper, K.A., et al., 2019. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer. Clinical Cancer Research 25, 1526–1534.

Cui, Y., Zhang, G., Liu, Z., Xiong, Z., Hu, J., 2018. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathological images. arXiv preprint arXiv:1803.02786 .

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., et al., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama 313, 1122–1132.

Graham, S., Rajpoot, N.M., 2018. Sams-net: Stain-aware multi-scale network for instance-based nuclei segmentation in histology images, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE. pp. 590–594.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A review. IEEE reviews in biomedical engineering 2, 147–171.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. ArXiv e-prints , arXiv:1703.06870arXiv:1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity Mappings in Deep Residual Networks. ArXiv e-prints , arXiv:1603.05027arXiv:1603.05027.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2016. Densely Connected Convolutional Networks. ArXiv e-prints , arXiv:1608.06993arXiv:1608.06993.

Javed, S., Fraz, M.M., Epstein, D., Snead, D., Rajpoot, N.M., 2018. Cellular community detection for tissue phenotyping in histology images, in: Computational Pathology and Ophthalmic Medical Image Analysis. Springer, pp. 120–129.

Khoshdeli, M., Parvin, B., 2018. Deep leaning models delineates multiple nuclear phenotypes in h&e stained histology sections. arXiv preprint arXiv:1802.04427 .

Kirillov, A., He, K., Girshick, R.B., Rother, C., Dollár, P., 2018. Panoptic segmentation. CoRR abs/1801.00868. URL: http://arxiv.org/abs/1801.00868, arXiv:1801.00868.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging 36, 1550–1560. doi:10.1109/TMI.2017.2677499.

Kwak, J.T., Hewitt, S.M., Xu, S., Pinto, P.A., Wood, B.J., 2015. Nucleus detection using gradient orientation information and linear least squares regression, in: Medical Imaging 2015: Digital Pathology, International Society for Optics and Photonics. p. 94200N.

LaTorre, A., Alonso-Nanclares, L., Muelas, S., Pea, J., DeFelipe, J., 2013. Segmentation of neuronal nuclei based on clump splitting and a two-step binarization of images. Expert Systems with Applications 40, 6521 – 6530. URL: http://www.sciencedirect.com/science/article/pii/S0957417413003904, doi:https://doi.org/10.1016/j.eswa.2013.06.010.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

Liao, M., qian Zhao, Y., hua Li, X., shan Dai, P., wen Xu, X., kai Zhang, J., ji Zou, B., 2016. Automatic segmentation for cell images based on bottleneck detection and ellipse fitting. Neurocomputing 173, 615 – 622. URL: http://www.sciencedirect.com/science/article/

pii/S0925231215011406, doi:https://doi.org/10.1016/j.neucom.2015.08.006.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Lu, C., Romo-Bucheli, D., Wang, X., Janowczyk, A., Ganesan, S., Gilmore, H., Rimm, D., Madabhushi, A., 2018. Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers. Laboratory Investigation 98, 1438.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis 33, 170 – 175. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301141, doi:https://doi.org/10.1016/j.media.2016.06.037. 20th anniversary of the Medical Image Analysis journal (MedIA).

Naylor, P., Laé, M., Reyal, F., Walter, T., 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging .

Nguyen, K., Jain, A.K., Sabata, B., 2011. Prostate cancer detection: Fusion of cytological and textural features. Journal of pathology informatics 2.

Raza, S.E.A., Cheung, L., Shaban, M., Graham, S., Epstein, D., Pelengaris, S., Khan, M., Rajpoot, N.M., 2018. Micro-Net: A unified model for segmentation of various objects in microscopy images. ArXiv e-prints , arXiv:1804.08145arXiv:1804.08145.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Sharma, H., Zerbe, N., Heim, D., Wienert, S., Behrens, H.M., Hellwich, O., Hufnagl, P., 2015. A multi-resolution approach for combining visual information using nuclei segmentation and classification in histopathological images., in: VISAPP (3), pp. 37–46.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Sirinukunwattana, K., e Ahmed Raza, S., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35, 1196–1206.

Sirinukunwattana, K., Snead, D., Epstein, D., Aftab, Z., Mujeeb, I., Tsang, Y.W., Cree, I., Rajpoot, N., 2018. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Scientific reports 8, 13692.

Veta, M., van Diest, P., Kornegoor, R., Huisman, A., Viergever, M., Pluim, J., 2013. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. PLoS ONE 8, e70221. doi:https://doi.org/10.1371/journal.pone.0070221.

Vu, Q.D., Graham, S., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kurc, T., Farahani, K., Zhao, T., et al., 2018. Methods for segmentation and classification of digital microscopy tissue images. arXiv preprint arXiv:1810.13230 .

Wang, P., Hu, X., Li, Y., Liu, Q., Zhu, X., 2016. Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing 122, 1–13.

Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., Dietel, M., Denkert, C., Klauschen, F., 2012. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. Scientific reports 2, 503.

Yang, X., Li, H., Zhou, X., 2006. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy. IEEE Transactions on Circuits and Systems I: Regular Papers 53, 2405–2414.

Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S., Chin, S.F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., et al., 2012. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Science translational medicine 4, 157ra143–157ra143.

Zhou, Y., Onder, O.F., Dou, Q., Tsougenis, E., Chen, H., Heng, P.A., 2019. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. arXiv preprint arXiv:1903.05358 .