DEGPR: Deep Guided Posterior Regularization for Multi-Class Cell Detection and Counting

Aayush Kumar Tyagi[1*], Chirag Mohapatra[1*], Prasenjit Das[3], Govind Makharia[3],
Lalita Mehra[3], Prathosh AP[2], Mausam[1]
[1]IIT Delhi [2]IISc, Bangalore [3]AIIMS, New Delhi
{tyagiaayushkumar, chirag131020, prasenaiims, govindmakharia, mehralalita9910, prathoshap}@gmail.com, mausam@cse.iitd.ac.in

Abstract

Multi-class cell detection and counting is an essential task for many pathological diagnoses. Manual counting is tedious and often leads to inter-observer variations among pathologists. While there exist multiple, general-purpose, deep learning-based object detection and counting methods, they may not readily transfer to detecting and counting cells in medical images, due to the limited data, presence of tiny overlapping objects, multiple cell types, severe class-imbalance, minute differences in size/shape of cells, etc.

In response, we propose guided posterior regularization (DEGPR)*, which assists an object detector by guiding it to exploit discriminative features among cells. The features may be pathologist-provided or inferred directly from visual data. We validate our model on two publicly available datasets (CoNSeP and MoNuSAC), and on MuCeD, a novel dataset that we contribute. MuCeD consists of 55 biopsy images of the human duodenum for predicting celiac disease. We perform extensive experimentation with three object detection baselines on three datasets to show that* DEGPR *is model-agnostic, and consistently improves baselines obtaining up to 9% (absolute) mAP gains.*

1. Introduction

Multi-class multi-cell detection and counting (MC2DC) is the problem of identifying and localizing bounding boxes for different cells, followed by counting of each cell class. MC2DC aids diagnosis of many clinical conditions. For example, CBC blood test counts red blood cells, white blood cells, and platelets, for diagnosing anemia, blood cancer, and infections [13, 31]. MC2DC over malignant tumor images helps assess the resistance and sensitivity of cancer treatments [9]. MC2DC over duodenum biopsies is needed to compute the ratio of counts of two cell types for diagnosing celiac disease [6]. Cell counting is a tedious process and often leads to significant inter-observer and intra-observer variations [4, 8]. This motivates the need for an AI system that can provide robust and reproducible predictions.

Standard object detection models such as Yolo [21], Faster-RCNN [35] and EfficientDet [44] have achieved state-of-the-art performance on various object detection settings. However, extending these to detecting cells in medical images poses several challenges. These include limited availability of annotated datasets, tiny objects of interest (cells) that may be overlapping, similarity in the appearance of different cell types, and skewed cell class distribution. Due to the non-trivial nature of the problem, MC2DC models may benefit from insights from trained pathologists, e.g., via discriminative attributes. For instance, in duodenum biopsies, intraepithelial lymphocytes (IELs) are structurally smaller, circular, and darker stained, whereas epithelial nuclei (ENs) are bigger, elongated, and lighter. A key challenge lies in incorporating these expert-insights within a detection model. A secondary issue is that such insights may not always be available or may be insufficient – this motivates additional data-driven features.

We propose a novel deep guided posterior regularization (DEGPR) framework. Posterior regularization (PR) is an auxiliary loss [12], which enforces that the posterior distribution of a predictor should mimic the data distribution for the given features. We call our method deep guided PR, since we apply it to deep neural models, and it is meant to formalize the clinical guidance given by pathologists. DEGPR incorporates PR over two types of features, which we term explicit and implicit features. Explicit features are introduced through direct guidance by expert pathologists. Implicit features are learned feature embeddings for each class, trained through a supervised contrastive loss [22]. Subsequently, both features are feed into a Gaussian Mixture Model (GMM). DEGPR constrains the distributions over the predicted features to follow that of the ground truth features, via a KL divergence loss between them.

We test the benefits of DEGPR over three base object detection models (Yolov5, Faster-RCNN, EfficientDet) on

[*] Equal contribution

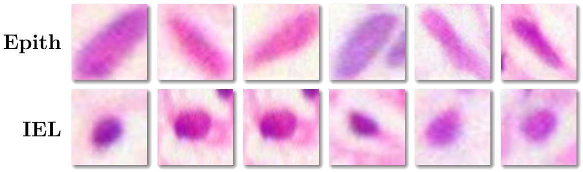

Figure 1. Visual dissimilarities between IELs and ENs. ENs (first row) are lighter stained, bigger and elongated in structure. IELs (second row) are darker stained, smaller, and circular in shape.

three MC2DC datasets. Of these, two are publicly available: CoNSeP [15] and MoNuSAC [47]. We additionally contribute a novel MuCeD dataset for the detection of celiac disease. MuCeD consists of 55 annotated biopsy images of the human duodenum, which have a total of 8,600 cell annotations of IELs and ENs. We find that DEGPR consistently improves detection and counting performance over all base models on all datasets. For example, on MuCeD, DEGPR obtains a 3-9% mAP gain for detection and a 10-35% reduction in mean absolute error for counting two cell types.

In summary, (a) we propose DEGPR to guide object detection models by exploiting the discriminative visual features between different classes of cells; (b) we use supervised contrastive learning to learn robust embeddings for different cell classes, which are then used as implicit features for DEGPR; (c) we introduce MuCeD, a dataset of human duodenum biopsies, which has 8,600 annotated cells of two types; and (d) we experiment on three datasets, including MuCeD, and find that DEGPR strongly improves detection and counting performance over three baselines. We release our dataset and code for further research.[*]

2. Related work

Object Detection in Medical Images: Object detection is the problem of localization and classification of objects of interest from an image. There are numerous object detection approaches in the literature, such as R-CNN [14], Yolo [34] and RetinaNet [27]. In this work, we experiment with Yolov5 – the latest in the Yolo series, Faster-RCNN – an improvement over R-CNN, and EfficientDet – the detection framework built on EfficientNet backbone [43].

There are two prominent ways of localization over medical images [28]. In cases where the exact location of an object is not required, detection is done by creating slices of images and subsequently performing classification on each patch [3, 11, 41]. In cases where location is important, standard object detection models are used after fine-tuning on medical datasets [19, 25, 26, 29, 36]. However, in most medical applications, limited availability of annotated data severely impacts the performance of fine-tuned models [42].

Methods for Cell Detection: One common approach for cell detection is to first perform object segmentation, followed by classification. Segmentation can provide a better solution for the detection task [10], as it is easier to impose spatial [1] or geometric [20, 45] priors over an explicit cell segmentation mask. At the same time, a pathologist's annotation effort in labeling segmentation masks is significantly higher than annotating bounding boxes. In the spirit of saving annotation effort, our work focuses on cell detection using annotated bounding boxes [5, 52]. An alternate annotation strategy to bounding boxes, is to annotate centroids [38, 49, 51] or use attention over feature maps [23, 39, 46]. It will be interesting to extend our work to these settings.

Object Detection for Cell Counting: Broadly, there are two main approaches proposed for cell counting in the literature: one is inspired by density-based methods, and the second models counting as a by-product of cell detection. Density-based methods use density maps instead of bounding boxes as labels and evade the hard task of localization [16, 24, 32]. Existing density-based approaches cannot handle multiple cell types, and hence cannot be directly used for our multi-class cell counting task. In the second approach, counting is generally done as a by product of predicted bounding boxes [2, 7]. It can also be done over predicted segmentation masks [30], but the challenge of data annotation for segmentation becomes relevant here too. DEGPR uses counting over predicted bounding boxes, and outperforms natural extensions of density based models.

3. Methods

In the problem of multi-class multi-cell detection and counting (MC2DC), we are given an input histopathology image im and the set of n different cell classes $C = \{c_1, c_2, \ldots c_n\}$. The goal is to output a set of bounding box sets $B = \{B_1, B_2, \ldots B_n\}$ where B_i denotes the set of output bounding boxes for the class c_i. These bounding boxes are then counted to obtain the count per cell class.

A possible solution for the aforementioned problem is an object detector D_θ (see Fig 2), which performs both bounding box detection ($B = D_\theta(im)$) and classification. Given ground-truth training data, object detectors are trained with a combination of objectness, classification, and localization losses. Objectness loss ($\mathcal{L}_{obj}$) is the confidence score indicating whether the box contains an object or not. Classification loss ($\mathcal{L}_{cls}$) is computed as cross-entropy between the predicted class and ground truth class. Localization loss ($\mathcal{L}_{loc}$) is the error in predicted bounding box coordinates as compared to ground truth bounding box coordinates. The total detection loss is given by Eq 1.

$$\mathcal{L}_{det} = \mathcal{L}_{obj} + \mathcal{L}_{cls} + \mathcal{L}_{loc} \tag{1}$$

While object detectors are feasible for MC2DC, in prac-

[*] https://github.com/dair-iitd/DeGPR

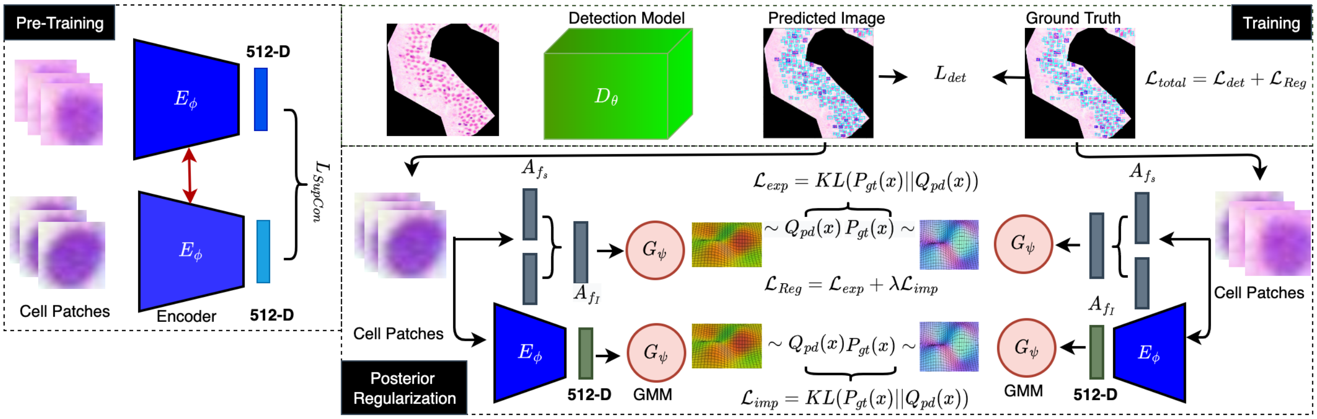

Figure 2. In pre-training stage cell patches are used to pre-train the contrastive encoder to differentiate the cell types. During training step, cell patches are compute based on predicted image from object detection model and ground truth images. These patches are used to compute average size (A_{f_s}), intensity (A_{f_I}) and contrastive embedding which will be used as feature vector to train GMM model (G_ψ). $P_{gt}(x)$ and $Q_{pd}(x)$ are sampled from GMM's and KL divergence is computed which is used to compute L_{reg}.

tice, they get bogged with issues such as the presence of limited annotated data, tiny objects, and multiple cell types. To address this, we propose a novel model architecture (DeGPR), which adds additional components and loss terms, and helps in training a more robust D_θ.

In particular, DeGPR utilizes *explicit* cell discriminative features (e.g., intensity and size for EN vs IEL) by comparing the distributions of these features over the ground truth and the predicted bounding boxes of each cell type. Additionally, DeGPR computes *implicit* feature embeddings for each bounding box, by training an encoder E_ϕ. It takes as input an image patch corresponding to a cropped out bounding box b to generate an embedding vector $E_\phi(b)$. Using a supervised contrastive (SupCon) loss for learning these embeddings ensures that they are well separated for different class types. It is to be noted that the explicit features are hand-crafted while the implicit features are data-driven – trained without any prior knowledge.

DeGPR uses both types of features to fit a Gaussian mixture model (GMM) defined by G_ψ for both ground truth and predicted bounding boxes. As shown in the Fig 2 (Posterior Regularization), it samples from the learned GMM model G_ψ and imposes similarity between the predicted and ground truth distributions via the Kullback-Leibler (KL) divergence loss between them – we call this the DeGPR loss. DeGPR loss is added to $\mathcal{L}_{det}$ and backpropagated to update the parameters θ. E_ϕ is pretrained using SupCon over gold bounding boxes, along with data augmentation and balanced subsampling of classes.

3.1. Deep Guided Posterior Regularization

DeGPR encourages discrimination between cell classes via differences in (explicit or implicit) features. Given a feature f_j, our method first computes the average feature value $(\mathcal{A}_{f_j}(c))$ for each class c, over the bounding boxes (B_c) of that class. For a pair of classes, DeGPR then computes the difference in these average values $(\mathcal{D}_{f_j})$. All feature differences are concatenated to form vectors $(\mathcal{D}_F)$, over which GMMs are fit. Finally, we use KL-divergence between the GMMs fits of the true and predicted bounding boxes.

Formally, let $F = \{f_1, f_2, \ldots f_m\}$ be a set of m features (implicit and explicit), where $f_j(b)$ denotes the value of j^{th} feature computed from a bounding box b. The average feature value of f_j for the class c is computed as:

$$\mathcal{A}_{f_j}(c) = \frac{1}{|B_c|} \sum_{b \in B_c} f_j(b) \tag{2}$$

Here, B_c is restricted to the bounding boxes for class c in a given image. For this image, the discriminative feature value $\mathcal{D}_{f_j}(c_i, c_k)$ for two classes c_i and c_k is defined as the difference of their average f_j values:

$$\mathcal{D}_{f_j}(c_i, c_k) = \mathcal{A}_{f_j}(c_i) - \mathcal{A}_{f_j}(c_k) \tag{3}$$

DeGPR concatenates the discriminative feature values corresponding to different features to form a discriminative feature vector denoted by $D_F(c_i, c_k)$:

$$\mathcal{D}_F(c_i, c_k) = [\mathcal{D}_{f_1}(c_i, c_k);\ \mathcal{D}_{f_2}(c_i, c_k); \ldots ; \mathcal{D}_{f_m}(c_i, c_k)] \tag{4}$$

Note, that each image in the dataset will have a discriminative feature vector corresponding to it. Once DeGPR has the set of discriminative feature vectors (D_F) for the entire minibatch, it learns the underlying feature distribution using a density estimator. We use Gaussian Mixture Models (GMM) for estimating the densities as they are known to be universal density approximators.

Two separate GMMs are learned for each of the ground truth and predicted bounding boxes/classes. That is, for every pair of classes c_i, c_k, we have a GMM P_{gt}, which models the discriminative feature vector $\mathcal{D}_F$ of the ground truth bounding boxes and another GMM Q_{pd}, similarly, for predicted bounding boxes. The goal is to 'align' these two feature vector distributions. DEGPR does this via a minimization of the KL divergence measure, given by Eq 5.

$$D_{KL}(P_{gt}||Q_{pd}) = \int_{\mathcal{X}} P_{gt}(x) \ln \frac{P_{gt}(x)}{Q_{pd}(x)} dx \tag{5}$$

Here, $\mathcal{X}$ represent the space of all features and $x \in \mathcal{X}$ are the individual feature vectors. DEGPR uses Monte Carlo estimates [18] to approximate the integral in Eq. 5 to get an estimate of the KL divergence using Eq 6. To do this, it treats the feature vectors obtained from the ground truth and predicted bounding boxes as samples of distributions P_{gt} and Q_{pd}, respectively. Let x_g and x_p be the ground and predicted feature vectors for image im, then KL-divergence is approximated as:

$$D_{MC} = \sum_{im} \log(P_{gt}(x_g)) - \log(Q_{pd}(x_p)) \tag{6}$$

Here, the sum is over all images im in the dataset. By the law of large numbers, D_{MC} converges to D_{KL} as number of samples → ∞ [18]. Hence, $D_{KL} \approx D_{MC}$.

DEGPR computes a loss term for each pair of classes c_i, c_k and then normalises them by the number of pairs, which is $\binom{n}{2}$. The final DEGPR loss $\mathcal{L}_F$ is calculated as:

$$\mathcal{L}_F = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} D_{KL}(P_{gt(i,k)}||Q_{pd(i,k)}) \tag{7}$$

We now describe our feature extraction process.

3.2. Explicit Features in DEGPR

Explicit features are hand-crafted to help discriminate between cell classes. Here, we use two such features, size (f_S) and intensity (f_I), nevertheless, DEGPR can also work with any other explicit features. Each explicit feature is modeled as a scalar (e.g., intensity has 1 scalar value).

Let (w_L, h_L) and (w_R, h_R) be the top left and bottom right pixel coordinates of a bounding box b, respectively. Then, we define the size ($f_S(b)$) of the bounding box as:

$$f_S(b) = (w_R - w_L) * (h_R - h_L) \tag{8}$$

Similarly, if $I(w,h)$ is the pixel intensity at (w,h), we define the intensity feature (f_I) as:

$$f_I(b) = \frac{\sum_{h=h_L}^{h_R} \sum_{w=w_L}^{w_R} I(w,h)}{f_S(b)} \tag{9}$$

With the size and intensity features computed as above, $\mathcal{D}_{f_I}(c_i, c_k)$ and $\mathcal{D}_{f_S}(c_i, c_k)$ from Eq 2 and 3 are used to obtain the explicit feature discriminative feature vectors:

$$\mathcal{D}_{F_{I,S}} = [\mathcal{D}_{f_I}(c_i, c_k);\ \mathcal{D}_{f_S}(c_i, c_k)] \tag{10}$$

Subsequently, we fit GMMs for $\mathcal{D}_{F_{I,S}}$ corresponding to the predicted and ground truth bounding boxes and compute the KL divergence between them, denoted by $\mathcal{L}_{\text{exp}}(c_i, c_k)$. The total explicit posterior regularization loss is given by adding these KL divergences for all pair of classes:

$$\mathcal{L}_{\text{exp}} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} \mathcal{L}_{\text{exp}}(c_i, c_k) \tag{11}$$

3.3. Implicit Features in DEGPR

The information provided by the experts (pathologists) about the discriminative features of the cell types may not be complete or may be hard to compute as an explicit feature. For instance, a shape-related feature (circular vs elongated) is hard to model, when exact segmentation masks are unavailable.* To deal with this, DEGPR adopts implicit feature learning and trains a ResNet18 [17] encoder E_ϕ, which converts an input image *patch* v to an implicit feature vector z_v. Here, each image patch corresponds to a predicted or ground truth bounding box (see Fig. 1). Since it may be difficult to learn a GMM on the ResNet18's 512-dimensional feature embedding z_v, DEGPR reduces it to a smaller size (10-22), using Principal Component Analysis (PCA), preserving 90% of explainable variance. The resulting features are denoted by F_{imp}.

Similar to the explicit features, DEGPR computes implicit feature discriminative vectors $\mathcal{D}_{F_{imp}}$ using Eq. 2 and 3. Subsequently, the KL divergence between the ground truth and predicted GMM fits of the implicit features, $\mathcal{L}_{imp}(c_i, c_k)$ is calculated for all class pairs, which are averaged to form the total implicit feature loss:

$$\mathcal{L}_{imp} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} \mathcal{L}_{imp}(c_i, c_k) \tag{12}$$

Pre-training of Feature Encoder: The encoder is pretrained using a supervised contrastive (SupCon) loss [22]. This operates on a pair of ground truth patches and penalizes the encoder if vectors for the patches of the same class are farther away in terms of a distance metric (such as Euclidean distance) compared to vectors for patches of different classes. SupCon is made hardness-aware by a temperature τ controling the strength of penalties on hard negative patches [48]. It is defined as follows:

$$L_{SupCon} = \sum_{v \in V} -\log \frac{1}{|P(v)|} \sum_{p \in P(v)} \frac{\exp(z_v z_p/\tau)}{\sum_a \exp(z_v z_a/\tau)} \tag{13}$$

[*] We tried edge detection over bounding boxes for computing shape features, but noise in highly zoomed medical images resulted in very poorly detected edges.

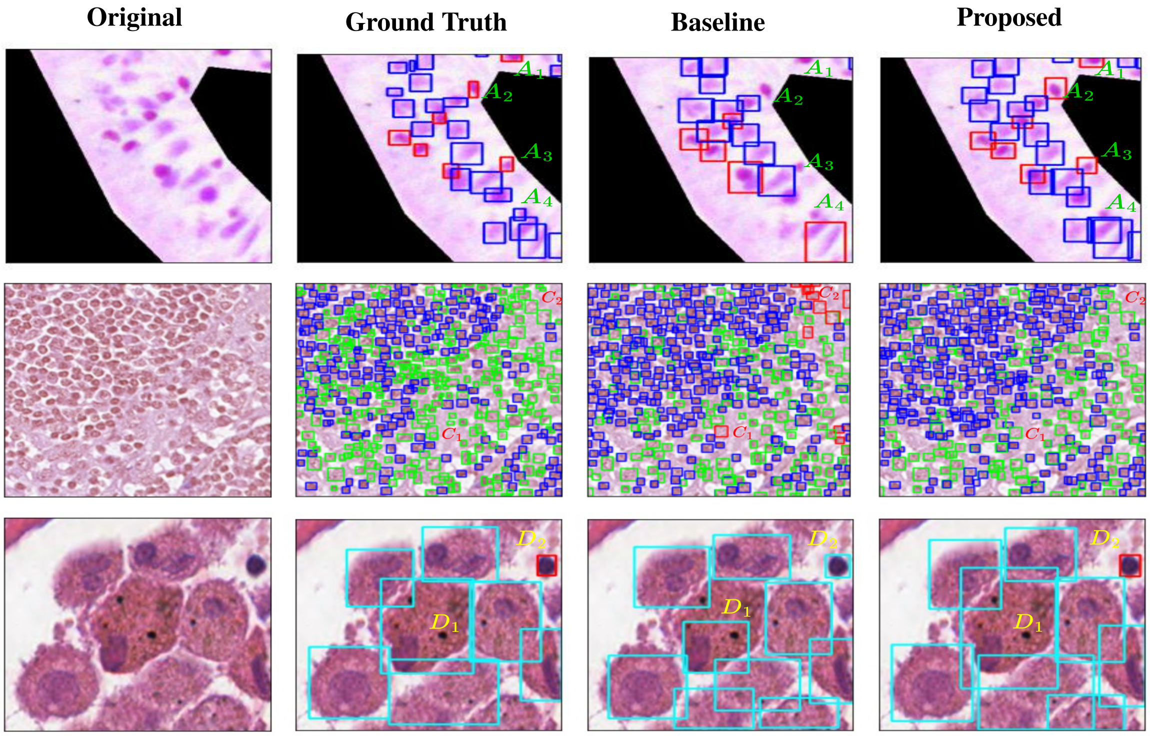

Figure 3. Qualitative performance of DEGPR. The third and fourth columns show Yolov5, with and without DEGPR. The first row is an image from MuCeD, with red and blue bounding boxes corresponding to IELs and ENs. The bounding boxes A_1, A_2, A_3 show improvement in detecting missing cells and A_4 shows improvements in misclassification. Row two is from the CoNSeP dataset with Inflammatory (red), Epithelial (blue), and Spindle (green) cells. Bounding boxes C_1 and C_2 show improvements in misclassification. Finally, the fourth row from the MoNuSAC dataset shows Epithelial (red), lymphocyte (blue), Neutrophil (green), and Macrophage (cyan) cells. D_1 shows improvement in bounding box prediction, while D_2 shows improvement in misclassification.

Here, L_{SupCon} is computed as a sum over set of all gold image patches V. For every patch v with a gold label c, a set of positive and negative pairs are defined as follows: p is a positive patch of v when it comes from a set $P(v)$, denoting other patches from class c. All the other patches $a \in V$, apart from p, form the negative samples.

With these, the objective of L_{SupCon} is to induce a representation space such that similar (positive) sample pairs are close to each other while dissimilar (negative) pairs are far apart. In our case, DEGPR applies L_{SupCon} on the supervised data, and thus the features learned help in discriminating between different cell-classes in the dataset.

Owing to the imbalance of cell classes in most of the datasets, we use balanced sampling of patches per class, when creating batches for training. Furthermore, since the predicted bounding boxes might not exactly overlap with the ground truth bounding boxes, for encoder robustness, we perform augmentation of the gold patches by randomly shifting and resizing the bounding boxes. Further, inspired by the idea of exposure bias [33] methods, we gradually introduce these augmented bounding boxes while training the encoder in an annealing manner. These approaches improve the performance of our encoder, and also of detector.

3.4. Loss Function

The detector D_θ is trained using a combination of object detection and posterior regularization losses. The latter is the sum of losses due to explicit and implicit features. We control the effect of regularization using λ_{reg}. We keep the encoder E_ϕ frozen when training the detector.

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda_{reg} \left(\mathcal{L}_{exp} + \mathcal{L}_{imp} \right) \tag{14}$$

4. Dataset Details

Multi-class Celiac Disease Dataset: We release MuCeD, a dataset that is carefully curated and validated by expert pathologists. The H&E-stained histopathology images of the human duodenum in MuCeD are captured through an Olympus BX50 microscope at 20× zoom using a DP26 camera with each image being 1920×2148 in dimension. The dataset has 55 images, with bounding boxes for 2,090 IELs and 6,518 ENs annotated using the LabelMe software and are further validated by multiple pathologists. These cells are selected from the epithelial area – a region of interest that has been explicitly segmented by experts. The epithelial area denotes the area of continuous villi and is used for cell detection, whereas rest of the area is masked out. Further, each image is sliced into 9 subimages and each subimage is re-scaled to 640x640, before it is given as input to object detection models. We divide 55 images into five folds of 11 images each and report 5-fold cross-validation numbers. Within 44 training images in a given fold, 8 are used for validation and 36 for training.

CoNSeP Dataset: To show the effectiveness of DEGPR, we further validate our method on two publicly available datasets. Colorectal nuclear segmentation and phenotypes (CoNSeP) [15] is a nuclear segmentation and classification dataset of H&E stained histology images. Each image is of 1000×1000 dimension and taken at 40× magnification. The dataset deals with single cancer, colorectal adenocarcinoma (CRA), images. It consists of a total of 41 whole slide images (WSI), which have a total of 24,319 annotated cells of 3 classes: inflammatory cells, epithelial cells, and spindle cells. A total of 27 images are used for training and the rest 14 images are used for testing purposes. Since CoNSeP is originally a segmentation dataset, to use it for MC2DC, we preprocess it by converting each segmentation mask into a bounding box. Further, we split the 1000×1000 images into 4 subimages of dimension 500×500. This results in an MC2DC dataset of 108 train and 56 test images.

MoNuSAC Dataset: We similarly use the multi-organ nuclei segmentation and classification (MoNuSAC) challenge dataset [47] by preprocessing segmentation masks into bounding boxes. MoNuSAC is a large dataset of nucleus boundary annotations and class labels. The dataset has over 46,000 nuclei from 37 hospitals, 71 patients, four organs, and four nuclei types. A total of 209 images (of 46 patients) are used for training and 85 images are used for testing. There are four nuclei types: epithelial nuclei, lymphocytes, neutrophils, and macrophages. Each cell type is different in structure and shape from the others. This makes the dataset perfect for our use case. The images are of variable size and we resize them to 640×640, for uniformity. Cells marked as ambiguous are filtered out from evaluation.

5. Experimental Setting

Through our experiments, we wish to answer the following research questions. (1) Is DEGPR model agnostic, i.e., can it be used effectively with multiple object detection models? (2) How much does DEGPR improve the cell detection and counting performance? And, (3) what are the incremental contributions of each of the various model components, such as implicit features, explicit features, and balanced training of the encoder?

Evaluation Metrics: We use precision, recall, and mean average precision (mAP) as the metrics for cell detection. For cell counting, we use MAE (mean absolute error) and MRE (mean relative error) as evaluation metrics. MAE provides the absolute difference between predicted count and true counts. MRE provides the relative difference with respect to the true counts. We compute MAE and MRE for the original complete image rather than subimages.

Additionally, we use the Q-histology [6] parameter for the quantitative classification of duodenum biopsy images into the celiac or non-celiac category. Q-Histology ratio is defined as the ratio of the number of IELs per 100 ENs. If the ratio is ≥ 25, then the patient suffers from celiac disease. We use this ratio to evaluate our model on the downstream task of classifying patients into celiac and non-celiac.

Baselines & Implementation Details: For MuCeD, we pretrain Yolov5 on the Kaggle data science bowl 2018 dataset,* which is a cell nuclei segmentation challenge, after converting the segmentation masks into bounding boxes. Yolov5 is trained for 300 epochs using SGD optimizer with a learning rate (lr) of 0.003, early stopping with patience 100 and batch size 32. We use Faster-RCNN with a Resnet50 backbone. We train Faster-RCNN for 200 epochs with the SGD optimizer and lr of 0.005, weight decay 0.0005 and lr scheduler with step size 3. Finally, for EfficientDet, we use the pretrained Efficientdet-d0 as the base model. We train EfficientDet for MuCeD with a lr of 0.001 for 2000 epochs with patience 100 and is trained with momentum optimizer [40]. For CoNSep and MoNuSac, EfficientDet is trained with a lr of 0.008, and Faster-RCNN and Yolov5 with 0.03 with lr scheduler with step size 3. All hyperparameters are fine-tuned using grid search on the respective validation sets. We conduct our experiments using NVIDIA-RTX 5000 and Tesla V100 GPUs.

While training DEGPR, we use 10^5 samples to approximate KL divergence in Eq 5 to get Eq 6. We perform a grid search to determine the best regularization factor λ_{reg} as $\lambda_{reg} = 0.01$ for MuCeD and CoNSep datasets, and $\lambda_{reg} = 0.001$ for MoNuSac. We also do grid search over relative weights of $\mathcal{L}_{exp}$ and $\mathcal{L}_{imp}$ and observe that 1:1 works best. All cell patches input to E_ϕ are cropped out from bounding boxes and resized to 224×224. Pretraining of ResNet18 encoder is done for 300 epochs with a lr of 0.001 and momentum [40] as an optimizer. For MuCeD dataset we observe that the model performs the best, when the IoU threshold is kept at 0.3. Hence, MuCeD experiments are performed for mAP:0.3. For CoNSeP and MoNuSAC, we use the standard IoU threshold of 0.5. We use horizontal flip, vertical flip, scaling and shifting as augmentation methods (more details in appendix).

Table 1. Detection and counting results for MuCeD

Model	Precision	Recall	mAP	MAE IEL	MRE IEL	MAE Epith	MRE Epith
Yolov5	0.711	0.723	0.751	8.97	42.62	14.61	13.43
Yolov5 (DeGPR)	**0.744**	**0.735**	**0.787**	**5.83**	**24.19**	**13.15**	**12.46**
Faster-RCNN	0.592	0.436	0.496	11.85	50.05	27.50	24.93
Faster-RCNN (DeGPR)	**0.646**	**0.468**	**0.541**	**9.61**	**31.64**	**26.50**	**23.60**
EfficientDet	0.266	0.640	0.414	20.35	133.91	20.30	20.78
EfficientDet (DeGPR)	**0.274**	**0.641**	**0.425**	**17.32**	**90.04**	**18.51**	**18.12**

Table 2. Detection and counting results for CoNSep

Model	Precision	Recall	mAP	MAE Inflm	MAE Epith	MAE Spindle	MAE Avg
Yolov5	0.638	0.574	0.606	28.21	55.50	57.93	47.21
Yolov5 (DeGPR)	**0.667**	**0.584**	**0.625**	**26.35**	**55.00**	**53.85**	**45.07**
Faster-RCNN	0.490	0.208	0.342	64.71	227.93	198.29	163.64
Faster-RCNN (DeGPR)	**0.571**	**0.331**	**0.451**	**51.93**	**151.28**	**163.00**	**122.07**
EfficientDet	0.633	0.178	0.205	86.00	79.86	134.36	100.27
EfficientDet (DeGPR)	**0.672**	**0.194**	**0.229**	**79.64**	**77.78**	**125.85**	**94.42**

Table 3. Detection and counting results for MoNuSac

Model	Precision	Recall	mAP	MAE-Epithelial	MAE-Lymphocyte	MAE-Neutrophil	MAE-Macrophage
Yolov5	0.611	**0.497**	0.481	25.15	14.12	1.96	3.95
Yolov5 (DeGPR)	**0.736**	0.474	**0.489**	**12.01**	**10.69**	**0.81**	**2.33**
Faster-RCNN	0.570	0.310	0.405	**19.52**	23.48	1.0	3.38
Faster-RCNN (DeGPR)	**0.643**	**0.370**	**0.473**	19.81	**22.44**	**0.82**	**3.02**
EfficientDet	0.256	**0.509**	0.402	17.67	17.25	1.24	6.51
EfficientDet (DeGPR)	**0.258**	0.499	**0.409**	**14.84**	**16.98**	**0.56**	**3.97**

Table 4. Ablation on MuCeD with Yolov5 baseline

Yolov5	explicit	implicit	Balance	Precision	Recall	mAP	MAE IEL	MRE IEL	MAE Epith	MRE Epith
✓				0.711	0.723	0.751	8.97	42.62	14.61	13.43
✓	✓			0.737	0.723	0.779	**5.79**	**23.13**	13.43	13.49
✓		✓		0.724	0.735	0.771	5.83	23.50	12.98	13.89
✓	✓	✓		0.741	0.734	0.780	6.06	23.40	**12.57**	13.05
✓	✓	✓	✓	**0.744**	**0.735**	**0.787**	5.83	24.19	13.15	**12.46**

6. Results

Detection and Counting Metrics: Table 1 shows the quantitative performance of DeGPR on MuCeD dataset. We compare the object detection models with and without DeGPR. We notice that there is a substantial improvement in all metrics and over all baselines, when DeGPR is used. It suggests that the guidance provided through explicit and implicit features helps the detection model to learn discriminating attributes for the cells. We particularly note that relative error for IEL counts (the minority class) has a drastic reductions of 18-43% points, showing the effectiveness of the approach. We observe a similar trend (tables 2 and 3) on CoNSeP and MoNuSAC datasets. While DeGPR performance is stronger than the baselines in all settings, we note that counting results in CoNSep are generally weak for all models – we suspect this is because the density of cells in that dataset is quite high (585 cells/image, compared to 70 and 156 for MoNuSac, MuCeD, resp.), and models end up missing a fraction of cells, leading to high absolute errors.

Qualitatively, we illustrate model predictions in Fig 3. The first column depicts the original image, the second is ground truth bounding boxes, the third shows the image with predictions from the Yolov5 baseline model, while the final column shows predictions from the Yolov5 with

[*] https://www.kaggle.com/c/data-science-bowl-2018

Table 5. Classification Metrics based on Q-Ratio

Measure	Baseline	Yolo+DEGPR
Recall	0.774	**0.936**
Precision	0.774	**0.871**
F1-score	0.774	**0.902**
Accuracy	0.746	**0.877**

Table 6. Counting vs Localization (MuCeD)

Model	MAE-IEL	MAE-Epith	MAE-Avg
UNet	11.72	26.85	19.29
FCRN-A	15.60	22.81	19.21
Countception	16.10	29.78	22.94
SAU-Net	11.56	28.07	19.82
Yolov5 (DEGPR)	**5.83**	**13.43**	**9.63**

Table 7. Counting vs Localization (CoNSeP)

Model	MAE-Inflamm	MAE-Epithelial	MAE-Spindle	MAE-Avg
UNet	64.03	101.11	159.47	108.20
FCRN-A	53.18	94.34	95.84	81.12
Countception	77.13	129.61	151.13	119.29
SAU-Net	50.72	77.38	99.14	75.75
Yolov5 (DEGPR)	**26.35**	**55.00**	**53.85**	**45.07**

DEGPR. Three rows contain an exemplar image each from MuCeD, CoNSeP, and MoNuSAC, respectively. We note that DEGPR reduces both misclassification and misidentification errors. The highlighted bounding boxes A_1, A_2, A_3 show improvement in the detecting missing cells, and A_4, C_1, C_2, D_2 show reductions in misclassification.

We further classify the patient samples in MuCeD, based on the Q-histology ratio. Table 5 reports the comparative analysis. DEGPR improves prediction accuracy from 74.55% to 87.7% and celiac F1-score from 0.774 to 0.902.

Comparison against Other Counting Models: Tables 6 and 7 compare the performance of counting via detection (Yolov5+DEGPR) vs density map based methods. For comparison, we use four state-of-the-art counting models: UNet [37], FRCN-A [50], Count-ception [32] and SAU-Net [16]. As all these methods expect a single-class input, we train separate models for each class, and aggregate. We observe that our approach outperforms all other methods by vast margins. Also, counting via detection in a multi-class setting is computationally convenient, as we can get the counts of all cell types from a single object detector. We also compare with MCSpatNet [1] and observe improved performance with DEGPR (see appendix).

Ablation Studies: We perform ablation analysis to understand the relative contributions of different components in the final performance. We run this study for MuCeD using Yolov5. All results are reported in Table 4. Comparing rows 1 and 2, we notice that explicit features improve precision from 0.711 to 0.737 and mAP from 0.751 to 0.774. We observe a similar performance gain in counting metrics. Similarly, introduction of implicit features (rows 1 vs. 3) improves mAP from 0.751 → 0.771 and MAE reduces from 8.79 → 5.83 along with improvements in other metrics. We also find that the implicit and explicit features capture complementary information (details in appendix Sec.5).

To mitigate class imbalance while contrastive pre-training of the encoder, we perform class-balanced sampling while creating batches, and additional data augmentations for robustness. Comparing rows 4 and 5, we note small improvements in most metrics, due to these.

Error Analysis: We also perform error analysis for our best model on MuCeD. The common failure modes include more errors in misclassifying IELs (minority class) as ENs than reverse. This is especially true if an IEL (circular) overlaps an EN (elongated), since the overall shape appears elongated. The darker stained images generally produce more errors, presumably because the intensity differences between cell types are diminished. Finally, EN cells are missed when they are very lightly stained.

7. Conclusions

We study multi-class cell detection and counting problems (MC2DC) over medical histopathological images in a limited data setting. Our solution, Deep Guided Posterior Regularization (DEGPR), imposes additional regularization terms, incentivizing the model to output, for each cell class, a posterior distribution of features over predicted bounding boxes similar to that of ground truth. DEGPR uses two types of features: explicit – generally provided by a domain expert, and implicit – trained automatically using a supervised contrastive loss over labeled data.

We also contribute a novel dataset of 55 duodenum biopsies (useful for predicting celiac disease) for our task, along with experimenting on two publicly available datasets. We find that DEGPR is effective in improving performance of several object detection backbones, obtaining substantial improvements in both detection and counting metrics. As a consequence, the F-score of the model in predicting celiac disease increases from 77% to 90%. We release our code and data for further research.

Acknowledgements: This work is primarily supported by IITD-AIIMS Multi-Institutional Faculty Interdisciplinary Research Project and an ICMR grant on automated grading of duodenal mucosal biopsies for celiac disease. It is also supported by Jai Gupta chair fellowship by IIT Delhi, Yardi School of AI for travel and honorarium grants. We thank the IIT Delhi HPC facility for computational resources.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4005–4014, 2021. 2, 8

[2] Carina Albuquerque, Leonardo Vanneschi, Roberto Henriques, Mauro Castelli, Vanda Póvoa, Rita Fior, and Nickolas Papanikolaou. Object detection for automatic cancer cell counting in zebrafish xenografts. *Plos one*, 16(11):e0260609, 2021. 2

[3] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013. 2

[4] Gino Roberto Corazza, Vincenzo Villanacci, Claudia Zambelli, Massimo Milione, Ombretta Luinetti, Carla Vindigni, Caterina Chioda, Luca Albarello, Daniela Bartolini, and Francesco Donato. Comparison of the interobserver reproducibility with different histologic criteria used in celiac disease. *Clinical Gastroenterology and Hepatology*, 5(7):838–843, 2007. 1

[5] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International conference on medical image computing and computer-assisted intervention*, pages 403–410. Springer, 2013. 2

[6] Prasenjit Das, Gaurav PS Gahlot, Alka Singh, Vandana Baloda, Ramakant Rawat, Anil K Verma, Gaurav Khanna, Maitrayee Roy, Archana George, Ashok Singh, et al. Quantitative histology-based classification system for assessment of the intestinal mucosal histological changes in patients with celiac disease. *Intestinal research*, 17(3):387, 2019. 1, 6

[7] Grzegorz Drałus, Damian Mazur, and Anna Czmil. Automatic detection and counting of blood cells in smear images using retinanet. *Entropy*, 23(11):1522, 2021. 2

[8] Arzu Ensari. Gluten-sensitive enteropathy (celiac disease): controversies in diagnosis and classification. *Archives of pathology & laboratory medicine*, 134(6):826–836, 2010. 1

[9] Maurizio Fazio and Leonard I Zon. Fishing for answers in precision cancer medicine. *Proceedings of the National Academy of Sciences*, 114(39):10306–10308, 2017. 1

[10] Seiya Fujita and Xian-Hua Han. Cell detection and segmentation in microscopy images with improved mask r-cnn. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[11] Jaime Gallego, Anibal Pedraza, Samuel Lopez, Georg Steiner, Lucia Gonzalez, Arvydas Laurinavicius, and Gloria Bueno. Glomerulus classification and detection based on convolutional neural networks. *Journal of Imaging*, 4(1):20, 2018. 2

[12] Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010. 1

[13] Beverly George-Gay and Katherine Parker. Understanding the complete blood count with differential. *Journal of Peri-Anesthesia Nursing*, 18(2):96–117, 2003. 1

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[15] 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. *Medical Image Analysis*, 58:101563, 2019. 2, 6

[16] Yue Guo, Oleh Krupa, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. Sau-net: A unified network for cell counting in 2d and 3d microscopy images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):1920–1932, 2022. 2, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[18] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007. 4

[19] Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *Machine Learning for Health Workshop*, pages 171–183. PMLR, 2020. 2

[20] Hao Jiang, Sen Li, Weihuang Liu, Hongjin Zheng, Jinghao Liu, and Yang Zhang. Geometry-aware cell detection with deep learning. *Msystems*, 5(1):e00840–19, 2020. 2

[21] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022. 1

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 1, 4

[23] Haijun Lei, Shaomin Liu, Ahmed Elazab, Xuehao Gong, and Baiying Lei. Attention-guided multi-branch convolutional neural network for mitosis detection from histopathological

images. *IEEE Journal of Biomedical and Health Informatics*, 25(2):358–370, 2020. 2

[24] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010. 2

[25] Chao Li, Xinggang Wang, Wenyu Liu, Longin Jan Latecki, Bo Wang, and Junzhou Huang. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Medical image analysis*, 53:165–178, 2019. 2

[26] Zhuoling Li, Minghui Dong, Shiping Wen, Xiang Hu, Pan Zhou, and Zhigang Zeng. Clu-cnns: Object detection for medical images. *Neurocomputing*, 350:53–59, 2019. 2

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[28] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 2

[29] Yang Liu, Zhuo Ma, Ximeng Liu, Siqi Ma, and Kui Ren. Privacy-preserving object detection for medical images with faster r-cnn. *IEEE Transactions on Information Forensics and Security*, 2019. 2

[30] Roberto Morelli, Luca Clissa, Roberto Amici, Matteo Cerri, Timna Hitrec, Marco Luppi, Lorenzo Rinaldi, Fabio Squarcio, and Antonio Zoccoli. Automating cell counting in fluorescent microscopy through deep learning with c-resunet. *Scientific Reports*, 11(1):1–11, 2021. 2

[31] Prashant Pandey, Vinay Kyatham, Deepak Mishra, Tathagato Rai Dastidar, et al. Target-independent domain adaptation for wbc classification using generative latent search. *IEEE Transactions on Medical Imaging*, 39(12):3979–3991, 2020. 1

[32] Joseph Paul Cohen, Genevieve Boucher, Craig A Glastonbury, Henry Z Lo, and Yoshua Bengio. Count-ception: Counting by fully convolutional redundant counting. In *Proceedings of the IEEE International conference on computer vision workshops*, pages 18–26, 2017. 2, 8

[33] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 5

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[36] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):1–7, 2018. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[38] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016. 2

[39] Tatsuhiko Sugimoto, Hiroaki Ito, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Multi-class cell detection using modified self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1855–1863, 2022. 2

[40] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 6

[41] Sairam Tabibu, PK Vinod, and CV Jawahar. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific reports*, 9(1):1–9, 2019. 2

[42] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 2

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[44] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 1

[45] Mohammad Tofighi, Tiantong Guo, Jairam KP Vanamala, and Vishal Monga. Prior information guided regularized deep learning for cell nucleus detection. *IEEE transactions on medical imaging*, 38(9):2047–2058, 2019. 2

[46] Naofumi Tomita, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA network open*, 2(11):e1914645–e1914645, 2019. 2

[47] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, and Amit Sethi. Multi-organ nuclei segmentation and classification challenge 2020. *IEEE transactions on medical imaging*, 39(1380-1391):8, 2020. 2, 6

[48] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. 4

[49] Thomas Wollmann and Karl Rohr. Deep consensus network: Aggregating predictions to improve object detection in microscopy images. *Medical Image Analysis*, 70:102019, 2021. 2

[50] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018. 8

[51] Yuanpu Xie, Fuyong Xing, Xiangfei Kong, Hai Su, and Lin Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 358–365. Springer, 2015. 2

[52] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2015. 2