ORIENTED CELL DETECTION AND CLASSIFICATION WITH TRANSFORMERS

Manel Davins Jovells

Universitat Politècnica de Catalunya
Introduction to Research

ABSTRACT

Automatic detection and classification of cell nuclei are important tasks in the medical field, particularly for clinical diagnostics, as they help pathologists perform rapid and accurate analysis of tissue samples. In this study, we propose a novel method to detect and classify cell nuclei using oriented bounding boxes by integrating image moments with detection transformers. Notably, our method does not require oriented bounding boxes as labels; instead, it leverages segmentation masks to extract the image moments, which are then used to derive the oriented bounding boxes. Incorporating cell orientation in the detection process is relevant because cells have diverse orientations, shapes, and sizes, overlapping boundaries, and tend to cluster densely, which requires precise localization for accurate analysis. By extracting image moments from segmentation masks, our model accurately detects and classifies cell nuclei with oriented bounding boxes. We validate our approach on the PanNuke dataset, demonstrating its effectiveness highlighting its potential for improving cell nuclei analysis in digital pathology.

Index Terms— Oriented Bounding Box Detection, Cell Detection, Image Moments

1. INTRODUCTION

1.1. Background

The integration of deep learning methods into digital pathology image analysis is transforming medical diagnostics by enabling more accurate and efficient analysis of Whole Slide Images (WSIs). Traditional methods for detecting and classifying cell nuclei often rely on segmentation techniques, which, despite their capacity to capture fine details, can be computationally intensive and less efficient, especially when processing giga-size WSIs. The clinically relevant tasks lie in the precise detection and classification of cells rather than segmentation. Object detection offers a more efficient approach by focusing on identifying and classifying objects within an image. In the context of digital pathology, object detection can be used to localize and classify cell nuclei, providing information for further analysis.

Standard bounding box detection methods, however, have limitations as they do not account for the orientation of objects. This is particularly important in cell detection, where nuclei are often irregularly shaped, oriented in various directions, and densely grouped. Traditional methods often result in overlapping bounding boxes, difficulting the differentiation of individual cells. On the other hand, oriented bounding box detection provides a more precise localization of the cell, leading to clearer and more understandable results.

1.2. Related work

Segmentation models such as HoVer-Net [1] and CellViT [2] have shown to be effective in the task of cell detection and classification. HoVer-Net consists of a U-Net-like architecture which contains three decoder branches: nuclear pixel (NP), horizontal-vertical (HV), and nuclear classification (NC). These branches predict the probability of a pixel belonging to a nucleus, the horizontal and vertical distances to the nucleus's center of mass, and the pixel label, respectively. A postprocessing step merges the outputs of the NP and HV branches to generate the final segmentation mask. CellViT extends HoVer-Net by replacing the convolutional encoder with a Vision Transformer (ViT) [3], achieving state-of-the-art performance, showcasing the effectiveness of transformers in medical image analysis.

Object detection is a fundamental task in computer vision, focusing on identifying and localizing objects within an image. Early methods like Region-based Convolutional Neural Networks (R-CNN) [4] have laid the foundation for more advanced techniques such as Fast R-CNN [5], Faster R-CNN [6] and YOLO (You Only Look Once) [7]. These methods typically involve generating region proposals and then classifying these proposals into categories.

The introduction of the Detection Transformer (DETR) [8] marked a significant advancement in object detection. DETR leverages transformers, originally designed for natural language processing, to capture global dependencies within an image. This approach removes the need for region proposals, simplifying the detection process and improving performance. Among different variants, Deformable-DETR [9] outperforms the original model by incorporating multi-scale deformable attention, which allow for more efficient and precise localization of objects of different sizes. These models have achieved state-of-the-art results on several object

detection benchmarks.

Based on Deformable-DETR, the Cell-DETR architecture [10] provides an efficient approach to cell detection and classification within WSIs. Cell-DETR consists of a hierarchical backbone that generates a multi-level feature pyramid for a given input image, followed by a multi-scale deformable transformer with 6 encoder and 6 decoder layers. The encoder enhances the input features through multi-scale deformable self-attention, while the decoder produces predictions for bounding boxes and labels based on a set of input object queries. This architecture supports the use of both ResNet-50 [11] and Swin Transformer [12] as backbones to extract input image features. By leveraging the capabilities of transformers, Cell-DETR efficiently processes large scale WSIs, achieving precise and scalable cell detection and classification.

Oriented bounding box detection extends standard bounding box detection by incorporating the orientation of objects. This approach is particularly relevant in applications where object orientation is important, such as aerial imagery, document analysis or digital pathology. Architectures like Rotated Region-based Convolutional Neural Networks (R-RCNN) [13] and Oriented Region Proposal Networks (O-RPN) [14] have been developed for this task. These methods achieve oriented object detection by predicting the angle of the bounding box in addition to the bounding box itself, allowing to generate bounding boxes that align with the object's true shape.

In addition to these methods, recent studies have explored the use of Gaussian distributions to model objects in the context of oriented bounding box detection. For instance, the Gaussian Wasserstein Distance (GWD) [15] and the Kullback-Leibler (KL) Divergence [16] have been employed to address the limitations of traditional regression losses. These approaches involve converting rotated bounding boxes into 2-D Gaussian distributions and calculating the divergence between these distributions to improve detection accuracy. This allows for better handling of objects with large aspect ratios and varying orientations, and provides a more informative loss even when there is no overlap between objects.

In this work, we adapt the architecture of Cell-DETR to work with image moments extracted from segmentation masks. Our method uses these predicted moments, which contain information about the shape and orientation of cell nuclei, to generate oriented bounding boxes. This allows us to capture detailed morphological information without requiring oriented bounding boxes as labels. We use the KL divergence loss to model cell nuclei as Gaussian distributions, allowing for more accurate representation and detection. By integrating image moments and probabilistic modeling, we enhance the precision of localization and classification, addressing the limitations of standard bounding box detection methods.

2. METHODOLOGY

2.1. Dataset

We use the PanNuke dataset to train and evaluate our model. This dataset contains a collection of images annotated for nuclei detection and classification. The PanNuke dataset [17] comprises 7,904 patches, each sized 256×256 pixels, extracted from WSIs in The Cancer Genome Atlas (TCGA) dataset, representing 19 diverse tissue types at a magnification of 40x. Within this dataset, there are 189,744 labeled nuclei cells categorized into five classes: neoplastic, inflammatory, connective, necrosis, and epithelial (see Figure 1).

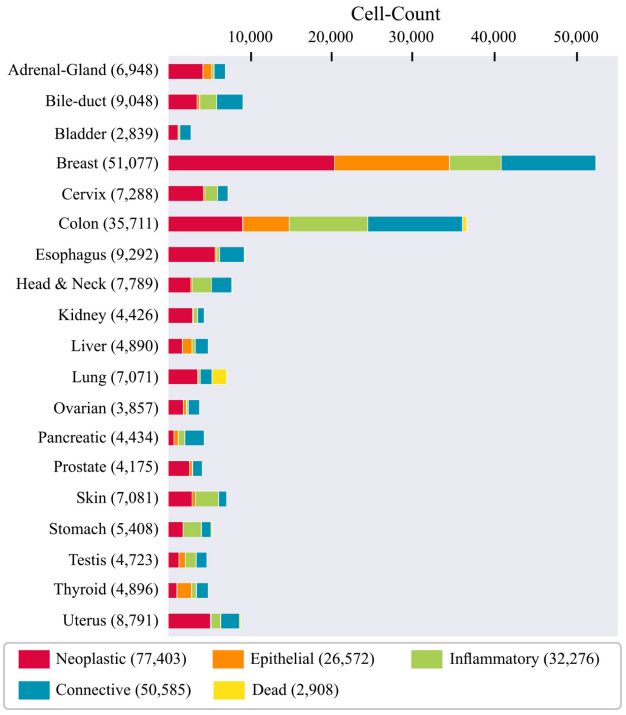

Fig. 1. PanNuke nuclei distribution overview for each of the nineteen tissue types, sorted by the total number of nuclei inside the tissue. The total number of nuclei within a tissue type is given in parentheses. Extracted from [2].

While the PanNuke dataset provides standard bounding boxes for the cells, it lacks oriented bounding box information. To address this, we leverage the cell segmentation masks and the cell topology to model each cell as an Gaussian distribution, extracting key image moments (centroid coordinates, covariance, and variances) that define the shape and orientation of each cell nuclei.

2.2. Image moments

Image moments are statistical properties of the pixel intensity distribution in an image and can be used to describe the shape and orientation of objects. In this work, we compute central

moments from the segmentation masks to capture features of each cell nucleus. The centroid coordinates ($\bar{x}$, $\bar{y}$) are given by the first-order moments:

$$\bar{x} = \frac{M_{10}}{M_{00}}, \quad \bar{y} = \frac{M_{01}}{M_{00}}$$

where M_{pq} are the spatial moments defined as:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y)$$

The central moments (μ_{11}, μ_{20}, and μ_{02}) are then calculated using the following formula:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)$$

where $I(x, y)$ represents the pixel intensity at position (x, y). These central moments are related to the covariance matrix of the object, which provides information about the object's shape and orientation. The covariance matrix is given by:

$$\Sigma = \begin{pmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{pmatrix}$$

The eigenvalues and eigenvectors of this covariance matrix correspond to the lengths and orientations of the major and minor axes of the ellipse that approximates the shape of the object. Specifically, we use these five moments to derive the parameters of the oriented bounding boxes, including the centroid (cx, cy), major and minor axis lengths (a, b), and the orientation angle (θ) as follows [18]:

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right)$$

$$a = \sqrt{2\left(\mu_{20} + \mu_{02} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}\right)}$$

$$b = \sqrt{2\left(\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}\right)}$$

2.3. Model architecture

This work builds upon the Cell-DETR architecture, which is designed for efficient cell detection and classification. Cell-DETR comprises a hierarchical backbone that generates a multi-level feature pyramid from an input image, followed by a multi-scale deformable transformer. The backbone can be either a ResNet-50 or a Swin Transformer pretrained on the COCO dataset. This backbone extracts features at different levels, which are then fed into the deformable transformer, which includes 6 encoder and 6 decoder layers.

To adapt Cell-DETR for oriented bounding box detection, we made minimal changes to the architecture. The main modification involved adapting the model to predict image moments instead of traditional bounding box coordinates. Specifically, we predict five moments: the centroid coordinates (cx, cy), the variances (μ_{20}, μ_{02}), and the covariance (μ_{11}). These predicted moments are then used to recover the angle and semi-axes of the ellipses, enabling the creation of rotated bounding boxes that accurately represent the shape and orientation of the nuclei, as described in the previous section.

2.4. Data preprocessing

In our preprocessing pipeline, the centroids (cx, cy) are normalized with respect to the image sides. The variances (μ_{20}, μ_{02}) and covariance (μ_{11}) are normalized using a min-max scaling approach based on the training set, removing outliers that distort the distribution. This ensures that the data fed into the model is standardized, facilitating more stable and effective training.

2.5. Data augmentation

Data augmentation is used to improve the model generalization capacity. Our augmentation pipeline includes elastic transformations, horizontal and vertical flips, 90-degree rotations, and blurring to simulate various image conditions. Additionally, we apply Hematoxylin-Eosin-DAB jitter to account for staining variability within the tissue sample.

2.6. Loss function

In this work, we replace the original Intersection over Union (IoU) loss used for bounding box prediction in Cell-DETR with a KL divergence loss. This choice is motivated by our modeling of each cell as a Gaussian distribution, defined by its centroid and covariance matrix extracted from the moments. The KL divergence loss measures the difference between the predicted and true Gaussian distributions in terms of both location and orientation. The moments are unnormalized before computing the divergence.

The KL divergence between two Gaussian distributions P and Q, with means μ_P and μ_Q and covariance matrices Σ_P and Σ_Q, is given by:

$$\begin{aligned} D_{KL}(P||Q) = \frac{1}{2}\Bigg[& \log \frac{|\Sigma_Q|}{|\Sigma_P|} - d \\ & + \text{tr}(\Sigma_Q^{-1}\Sigma_P) \\ & + (\mu_Q - \mu_P)^T \Sigma_Q^{-1} (\mu_Q - \mu_P) \Bigg] \end{aligned}$$

where d is the dimensionality of the distribution.

In addition to the KL divergence loss, we also employ an L1 loss for moment regression and a Cross-Entropy (CE) loss for the classification task. The L1 loss ensures accurate

regression of the moments and the CE loss optimizes the classification.

Similar to the original bounding box loss in DETR, our moment loss is defined as:

$$L_{\text{moment}}(m_i, \hat{m}_{\sigma(i)}) = \lambda_{\text{KL}} L_{\text{KL}}(m_i, \hat{m}_{\sigma(i)}) + \lambda_{\text{L1}} ||m_i - \hat{m}_{\sigma(i)}||_1$$

where λ_{KL} and λ_{L1} are hyperparameters, and m_i and $\hat{m}_{\sigma(i)}$ are the predicted and ground truth moments, respectively. Both losses are normalized by the number of objects inside the batch. This approach allows for the precise detection of the cells, taking into account their location and orientation.

3. EXPERIMENTS AND RESULTS

3.1. Implementation details

Our model is implemented using PyTorch and has been trained on a two NVIDIA GeForce RTX 3090 GPU for 100 epochs with a batch size of 2. The backbone of the network is a Swin Transformer pre-trained on the COCO dataset. For optimization, we use the AdamW optimizer with a learning rate of $2 \cdot 10^{-4}$ and weight decay of 10^{-4}. Other important hyperparameters include the costs of the moment loss, which are set the weight for the L1 loss term to 2 and for the KL divergence loss term to 4.

3.2. Experiments

We evaluate our model using the PanNuke dataset, which contains 7,904 image patches of size 256×256 from 19 different tissue types, annotated with five classes of nuclei. The dataset is split into training, validation, and test sets, ensuring representation of all tissue types and classes.

As quantitative results, we provide precision, recall, and F1-score for the detection task and for each class, computed on the test set. Additionally, we show the evolution of the Rotated IoU metric over the validation set to illustrate the improvement over time. Furthermore, we present visual results to demonstrate the qualitative performance of our model.

3.3. Results

In Table 1 we summarize the results of the experiments. Our method achieves a very similar performance in terms of classification compared to the baseline Cell-DETR model. We note that classes which are less frequent, such as necrosis, which has minimal representation in the dataset, tend to achieve lower results.

To further demonstrate the improvement in terms of oriented detection, we present a plot of the Rotated IoU metric over the validation set across epochs, which allows to understand the consistent improvement of the model in terms of detecting the orientation and shape of cell nuclei.

Metric	Precision	Recall	F1-score
Cell-DETR			
Detection	0.801	0.722	0.759
Neoplastic	0.705	0.642	0.672
Inflammatory	0.591	0.586	0.589
Connective	0.590	0.479	0.529
Necrosis	0.479	0.316	0.381
Epithelial	0.670	0.690	0.679
Ours			
Detection	0.785	0.770	0.777
Neoplastic	0.707	0.668	0.687
Inflammatory	0.550	0.570	0.560
Connective	0.538	0.535	0.536
Necrosis	0.375	0.265	0.311
Epithelial	0.676	0.702	0.689

Table 1. Comparison of precision, recall, and F1-score for detection and each cell class between our model and Cell-DETR.

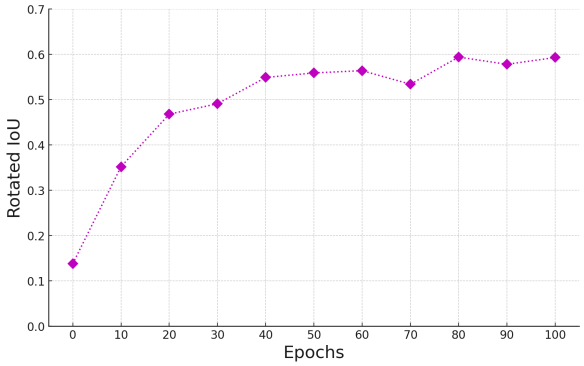

Fig. 2. Evolution of the rotated intersection over union metric on the validation set over the training epochs for our model.

We also provide visual results to illustrate the performance of our method. Figure 3 shows sample images with the predicted bounding boxes overlaid, comparing the predicted oriented bounding boxes with the ground truth.

4. CONCLUSIONS

In this report, we have presented a novel approach for detecting and classifying cell nuclei using oriented bounding boxes by integrating image moments to the Cell-DETR model. By leveraging existing segmentation masks as labels, our method does not require oriented bounding boxes, which simplifies the labeling process and broadens the applicability of the model.

Quantitative results indicate that the performance of our model is comparable to the original Cell-DETR in terms of classification accuracy and detection. Visual results demonstrate that the localization of the bounding boxes generated

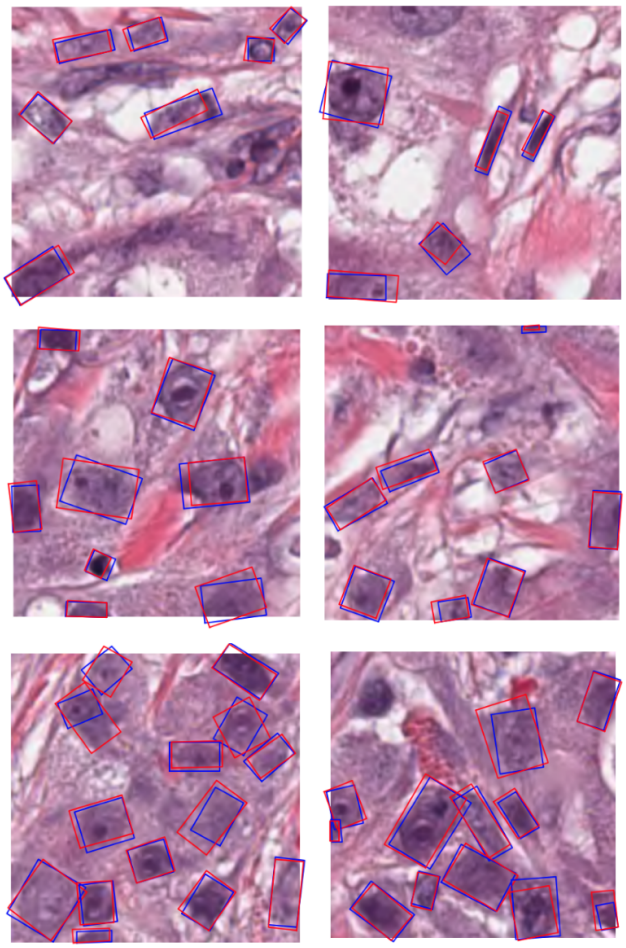

Fig. 3. Examples with predicted oriented bounding boxes in blue and ground truth oriented bounding boxes in red.

by our method is accurate and clear, particularly in situations with dense clustering of cells. This improvement shows great potential for applications in digital pathology, where accurate and detailed cell localization is important.

Future work will focus on continuing with the development of the model by adapting key features of the original Deformable DETR model, such as the two-stage DETR and the iterative refinement mechanism. These features have shown to improve model performance and could further increase the effectiveness of our approach in detecting and classifying cell nuclei.

5. REFERENCES

[1] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Jin Tae Kwak, and Nasir M. Rajpoot, "XY network for nuclear segmentation in multi-tissue histology images," *CoRR*, vol. abs/1812.06499, 2018.

[2] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek, "Cellvit: Vision transformers for precise cell segmentation and classification," 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1–1, 12 2015.

[5] Ross B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015.

[6] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[7] Ross Girshick Ali Farhadi Joseph Redmon, Santosh Divvala, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020.

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable DETR: deformable transformers for end-to-end object detection," *CoRR*, vol. abs/2010.04159, 2020.

[10] Oscar Pina, Eduard Dorca, and Verónica Vilaplana, "Cell-detr: Efficient cell detection and classification in wsis with transformers," in *Proceedings of Machine Learning Research*. 2024, pp. 1–14, MIDL.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[13] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang, "Rotated region based cnn for ship detection," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 900–904.

[14] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han, "Oriented R-CNN for object detection," *CoRR*, vol. abs/2108.05699, 2021.

[15] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian, "Rethinking rotated object detection with gaussian wasserstein distance loss," *CoRR*, vol. abs/2101.11952, 2021.

[16] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan, "Learning high-precision bounding box for rotated object detection via kullback-leibler divergence," *CoRR*, vol. abs/2106.01883, 2021.

[17] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot, "Pannuke dataset extension, insights and baselines," 2020.

[18] François Chaumette, "Image moments: A general and useful set of features for visual servoing," *Robotics, IEEE Transactions on*, vol. 20, pp. 713 – 723, 09 2004.