

# AI-assisted prostate cancer detection and localisation on biparametric MR by classifying radiologist-positives

Xiangcen Wu<sup>a</sup>, Yipei Wang<sup>a</sup>, Qianye Yang<sup>a</sup>, Natasha Thorley<sup>b</sup>, Shonit Punwani<sup>b</sup>, Veeru Kasivisvanathan<sup>b</sup>, Ester Bonmati<sup>a,c</sup>, and Yipeng Hu<sup>a</sup>

<sup>a</sup>Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK

<sup>b</sup>Div of Surgery & Interventional Sci, University College London, London, UK

<sup>c</sup>School of Computer Science and Engineering, University of Westminster, London, UK

## ABSTRACT

Prostate cancer diagnosis through MR imaging have currently relied on radiologists' interpretation, whilst modern AI-based methods have been developed to detect clinically significant cancers independent of radiologists. In this study, we propose to develop deep learning models that improve the overall cancer diagnostic accuracy, by classifying radiologist-identified patients or lesions (i.e. radiologist-positives), as opposed to the existing models that are trained to discriminate over all patients. We develop a single voxel-level classification model, with a simple percentage threshold to determine positive cases, at levels of lesions, Barzell-zones and patients. Based on the presented experiments from two clinical data sets, consisting of histopathology-labelled MR images from more than 800 and 500 patients in the respective UCLA and UCL PROMIS studies, we show that the proposed strategy can improve the diagnostic accuracy, by augmenting the radiologist reading of the MR imaging. Among varying definition of clinical significance, the proposed strategy, for example, achieved a specificity of 44.1% (with AI assistance) from 36.3% (by radiologists alone), at a controlled sensitivity of 80.0% on the publicly available UCLA data set. This provides measurable clinical values in a range of applications such as reducing unnecessary biopsies, lowering cost in cancer screening and quantifying risk in therapies.

**Keywords:** Prostate Cancer, Prostate biopsy, Deep Learning, Image Classification

## 1. INTRODUCTION

Multi-parametric MR (mpMR) imaging has been developed for a non-invasive alternative to previous blind biopsies, for detecting and localise clinically significant prostate cancer (csPC). This has motivated many machine learning (ML) methods developed for this MR-based radiological task. For classifying patients with csPC from those who may safely avoid unnecessary biopsy confirmation, studies showed that ML models achieved as accurately as radiologists,<sup>1</sup> or even better.<sup>2</sup> However, both radiologists and ML classifiers are still subject to limited performance. For example, the PROMIS study showed a specificity of 38.9% at a controlled sensitivity of 90%. Few studies evaluated the diagnostic accuracy at local level, such as localising individual lesions or Barzell zones<sup>3,4</sup> - by radiologists or ML models despite its clinical utilities in targeted biopsy, therapy planning and evaluation. This is in part due to a lack of data with comprehensive histopathology labels, for robust validation. Efforts have been made, however, through the use of template-based saturation biopsy, e.g. in the PROMIS study,<sup>5</sup> and a mixed targeted and systematic biopsy,<sup>6</sup> to provide better sampling of the disease.

The existing ML models, which have access to both radiologist and histopathology labels during training, has the potential to improve the radiologist-only performance.<sup>7</sup> Histopathology labels obtained from a mixed targeted and systematic biopsy (e.g. the UCLA data set<sup>6</sup> used in this study) are more feasible, therefore more available, compared with saturation biopsy. However, radiologist-negative regions are sampled randomly and highly sparsely (e.g. 3 - 30 needle biopsied specimens) in the former. This directly leads to a highly variable and, potentially, unknown biased histopathology labels (biased due to, e.g., cohort, observer, imaging protocol). Our preliminary experience suggested that training ML models learned from such labels can be challenging and inefficient. In contrast, the radiologist-positive cases are sampled by targeted biopsy which yields lower (albeit perhaps still significant) variance and bias, which motivated this work.

In this work, we aim to develop an AI system that provides assistance to radiologist MR reading, rather than performing independently of the radiologists. We propose to train the AI models solely on classifying radiologist-positive cases (at lesion, Barzell zone and patient levels), to avoid using the “radiologist-negative cases” with the above-discussed label issues. We argue that those radiologist-positive cases have 1) a better-defined population distribution (based on continuously improving radiology guidelines) and 2) better sampled histopathology labels (e.g. through targeted biopsy), compared to the “entire” population (i.e. with radiologist-negative cases) often with varying entry criteria and lacking consistent pathology labels from often under-sampled patients or regions. The developed classifiers then can be combined with the radiologists to provide the final and improved diagnosis.

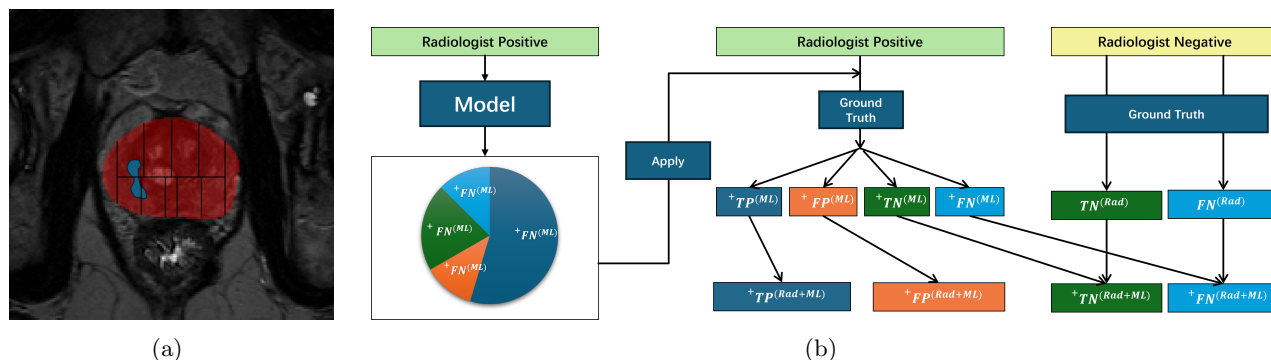


Figure 1: (a) A suspicious lesion covering three zones based on Barzell zone template. (b) A diagram illustrates the proposed classification of radiologist-positive cases.

## 2. METHOD

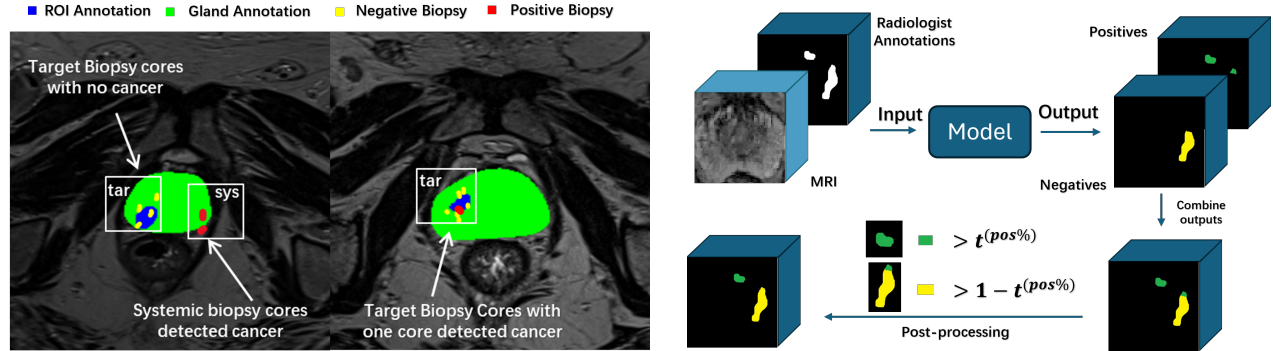
### 2.1 Classification of radiologist-positive cases

Training an ML model to classify MR voxels, histopathology ground-truth is required at every voxel, to compute the number of true positive  $TP^{(ML)}$ , false positive  $FP^{(ML)}$ , true negative  $TN^{(ML)}$  and false negative  $FN^{(ML)}$  cases. Similarly, the radiologist performance can be measured by  $TP^{(Rad)}$ ,  $FP^{(Rad)}$ ,  $TN^{(Rad)}$  and  $FN^{(Rad)}$ . In targeted biopsy, annotation from radiologists indicates suspicious regions of interest (ROIs) which undergo needle-based biopsy sampling for subsequent histopathology examination. Given a definition of clinically significant cancer (e.g. Gleason group  $\geq 3+4$ ), our proposed ML models classify those radiologist-positive ROIs ( $+ROIs$ ) into  $+TP^{(ML)}$ ,  $+FP^{(ML)}$ ,  $+TN^{(ML)}$  and  $+FN^{(ML)}$ , based on positive/negative pathology. The prefix  $+$  indicates the classification of radiologist-positive cases, where the “cases” refers to radiologist’s annotation on MR imaging voxels during model training. The voxels outside of these  $+ROIs$  are, by definition, considered as negative (or a separate background class) labels during training, therefore their histopathology labels are not required. During model evaluation, these classified voxels are subsequently grouped and analysed at different sampling levels (i.e. now “cases” are lesions, Barzell zones<sup>3</sup> or patients, based on available ground-truth data). The following section describes our proposed implementation, using the radiologist-positive cancer classification system, to classify all cases into  $TP^{(Rad+ML)}$ ,  $FP^{(Rad+ML)}$ ,  $TN^{(Rad+ML)}$  and  $FN^{(Rad+ML)}$ , indicated in Figure 1b. It is noteworthy that  $+TP^{(ML)} = TP^{(Rad+ML)}$  and  $+FP^{(ML)} = FP^{(Rad+ML)}$ .

### 2.2 Segmentation networks

A segmentation network is adopted to take MR images and the radiologist annotation as the multi-channel network input. In this work, a two-channel input contains a T2-weighted (T2w) MR volume and corresponding radiologist-positive segmentation mask, and a four-channel bi-parametric MR (bpMR) input uses a T2w, a diffusion-weighted with high b-value ( $DWI_{hb}$ ), an ADC map and a segmentation mask. The input may be configured to take further imaging modalities or, potentially, other clinical data if available. Although mpMR imaging is now widely accepted for accurate prostate detection, testing the case of using uni-modal T2w-only input is interesting. This is because classifying radiologist-positive cases has not been previously explored, thus the sensitivity of the learned model discriminating ability to different modalities is yet investigated. The segmentation model is trained to predict a 3-class class probability map, i.e. with three output channels representing

the positive, negative and background classes, supervised by the radiologist-positive labels, described in Secs. 2.1 and 3.2. During inference, a threshold is used in this work, for adjustment between Type I and Type II errors in subsequent classification tasks. The threshold  $t^{(pos\%)} \in [0, 1]$  indicates the percentage of positive voxels in a ROI, over which the radiologist-defined ROIs are classified as positive/negative ROIs, based on which voxel class dominates the ROI, as illustrated in Fig. 2b.



(a) Radiologist true (right) and false (left) positive  
Figure 2: Determination of suspicious ROIs as showed on (a), and (b) structural diagram of the deep learning prediction and classification of positive/negative on radiologist's annotations.

### 3. EXPERIMENTS

#### 3.1 Data sets labelled by varying biopsy protocols and network training

The open UCLA data set<sup>6</sup> are used primarily for model training, with T2w, ADC, and  $DWI_{hb}$  sequences. 898 and 825 imaging studies, for respective T2w-only and bpMR, are annotated with both radiologist ROIs (with 0-5 UCLA scores) and histopathology ground-truth. The ground-truth is based on a mixed targeted biopsy and systematic random biopsy outside of the targeted ROIs. All image volumes are center-cropped based on the radiologist annotation of the prostate gland, intensity-normalized, and resampled to an image size of  $64 \times 64 \times 64$ . The recently published PROMIS data set<sup>5</sup> are used in an additional validation of the radiologist-and-ML combined performance, in part due to its more densely sampled histopathology results outside of the radiologist-positive ROIs (with 1-5 PIRADS scores), based on Barzell zones.<sup>3</sup>

We utilize SwinUNETR-v2<sup>8</sup> as our segmentation model. All results reported in this study are based on multiple random train-test splits, repeated and averaged over five times. For each split, we train the model for a fixed 100 epochs using the AdamW optimizer with a learning rate of  $5e^{-5}$ . During training, we apply random affine transformations, Gaussian smoothing, Gaussian noise, and contrast adjustment, each with a probability of 0.25.

#### 3.2 Patient, ROI and Barzell zone classification

The results obtained with the ML model were compared with the decision of the radiologist on three levels. Varying cutoff values  $\alpha$  for defining significant cancer with respect to the UCLA score has been tested  $\alpha \in [2, 5]$ . A visual explanation of the comparison between the AI model and the radiologist is showed on Fig. 3.

##### 3.2.1 Patient Level classification

Patient-level accuracy can be computed on the UCLA data set. A positive patient is indicated by the presence of any positive biopsy. For the radiologist's decision, ROIs have been classified at different UCLA score cutoffs  $\in [2, 5]$ . For AI models, by varying the threshold  $t^{(pos\%)}$ , the radiologist-and-ML combined sensitivity  $sen^{(Rad+ML)}$  and specificity  $spc^{(Rad+ML)}$  are compared with the radiologist performance  $sen^{(Rad)}$  and  $spc^{(Rad)}$ .

### 3.2.2 ROI Level classification

For classifying the radiologist-positive ROIs, the ML model performance,  $+sen^{(ML)}$  and  $+spc^{(Rad)}$ , is based on  $+TP^{(ML)}$ ,  $+FP^{(ML)}$ ,  $+TN^{(ML)}$  and  $+FN^{(ML)}$  (Sec. 2.1). As this work aims to classify the positive ROIs deemed by radiologist examination, the  $+sen^{(ML)}$  and  $+spc^{(Rad)}$  are based on radiologist's ROIs only. All radiologist's ROIs have been classified by both radiologist and AI models.

### 3.2.3 Zone Level classification

For assessing localisation performance taking into account radiologist-negative cases, we classify Barzell zones into different csPC definitions as defined in the PROMIS study, with varying PIRADS cutoffs, to compare the resulting radiologist-and-ML combined  $sen^{(Rad+ML)}$  and  $spc^{(Rad+ML)}$  and the radiologist performance  $sen^{(Rad)}$  and  $spc^{(Rad)}$ . The radiologist performance can be obtained from the PROMIS data set, based on the Barzell-zone-level  $TP^{(Rad)}$ ,  $FP^{(Rad)}$ ,  $TN^{(Rad)}$  and  $FN^{(Rad)}$ . The developed ML model is applied on the radiologist-positive Barzell zones, see Fig 1a, based on the same approach described in Sec. 2.2, on the UCLA test data. This obtained radiologist-positive zone classification is then assumed on the PROMIS data set to modify the radiologist classification, to obtain a new set of  $TP^{(Rad+ML)}$ ,  $FP^{(Rad+ML)}$ ,  $TN^{(Rad+ML)}$  and  $FN^{(Rad+ML)}$ , as described in Sec. 2.1.

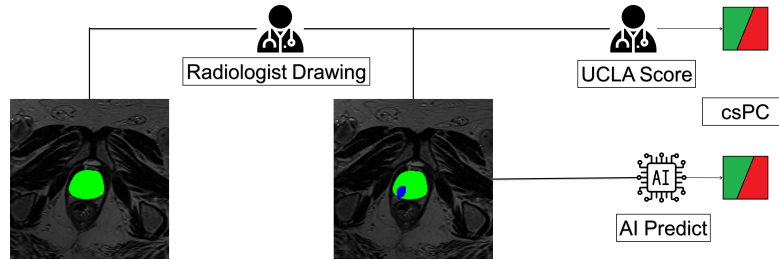


Figure 3: Radiologists first mark ROI and assign a UCLA score. Then, varying cutoffs are tested for determining whether the lesion is significant. The AI model, takes the MRI and the ROI as input and predicts the likelihood of csPC.

## 4. RESULTS AND DISCUSSION

The ROI-level, zone-level and patient-level results are summarised in Table. 1. The results are reported as example sensitivity and specificity, at varying diagnostic cutoffs and csPC definitions, for assessing their respective clinical values and a direct comparison between methods, as detailed in Sec. 3.2. The overall accuracy metrics, including average precision and area under the ROC curves, are excluded because their clinical relevance is unclear. As seen from Table. 1a, the high  $+sen$  and  $+spc$  in classifying  $+ROIs$  are obtained at varying radiologist score cutoffs, for both T2w-only and bpMR cases, with visual examples provided in Fig. 4 that with a higher UCLA score, model has more confidence predicting the ROI as a positive lesion. For classifying Barzell zones, a  $spc^{(Rad+ML)}$  of 90% was obtained by adding the proposed ML models, improved from the radiologists'  $spc^{(Rad)}$  of 72.5%, at the same  $sen^{(Rad+ML)} = sen^{(Rad)}$ , as in Table. 1b. The improvement was also readily observed at the patient classification, as shown in Table. 1c.

Our work demonstrate a wide applicability of the proposed radiologist-positive classification approaches, for both patient risk stratification and csPC localisation, as well as in different clinical scenarios, such as screening that requires high sensitivity and managing high-risk procedures that may benefit from high specificity. In practice, it can be utilized during the radiologist's annotation process to provide a confidence score to the annotator or used independently, potentially eliminating the need for further biopsy.

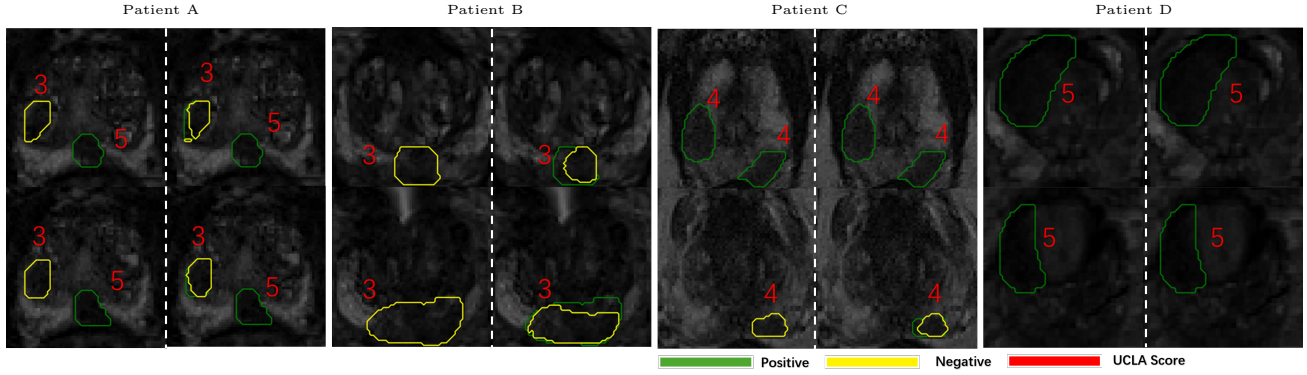


Figure 4: Four example patients were utilized here on the figure. Left side of the dashed line represents the ground truth data, while the right side displays the model's predictions.

|                                        | Sen    | Spe    |
|----------------------------------------|--------|--------|
| <b>UCLA <math>\geq 3</math> (T2)</b>   | 94.20% | 97.14% |
| <b>UCLA <math>\geq 5</math> (T2)</b>   | 92.59% | 76.42% |
| <b>UCLA <math>\geq 3</math> (bpMR)</b> | 76.92% | 95.41% |
| <b>UCLA <math>\geq 5</math> (bpMR)</b> | 91.30% | 52.27% |

(a) ROI level results on  $^+ROIs$

|                             | Sen           | Spe           |
|-----------------------------|---------------|---------------|
| <b>Promise Radiologist</b>  | <b>58.66%</b> | <b>72.53%</b> |
| $t^{(pos\%)} = 0.5$ (T2)    | 38.20%        | 92.18%        |
| $t^{(pos\%)} = 0.3$ (T2)    | 39.29%        | 91.80%        |
| $t^{(pos\%)} = 0.01$ (T2)   | 45.69%        | 91.10%        |
| $t^{(pos\%)} = NaN$ (T2)    | <b>58.66%</b> | <b>90.42%</b> |
| $t^{(pos\%)} = 0.5$ (bpMR)  | 30.09%        | 93.10%        |
| $t^{(pos\%)} = 0.3$ (bpMR)  | 34.11%        | 92.58%        |
| $t^{(pos\%)} = 0.01$ (bpMR) | 37.39%        | 90.56%        |
| $t^{(pos\%)} = NaN$ (bpMR)  | <b>58.66%</b> | <b>75.70%</b> |

(b) Zone level results on PROMIS

|                                    | Sen           | Spe           |
|------------------------------------|---------------|---------------|
| <b>*UCLA <math>\geq 2</math></b>   | 92.24%        | 6.44%         |
| <b>*UCLA <math>\geq 3</math></b>   | 92.19%        | 7.84%         |
| <b>*UCLA <math>\geq NaN</math></b> | <b>80.00%</b> | <b>36.33%</b> |
| <b>*UCLA <math>\geq 4</math></b>   | 65.47%        | 70.30%        |
| <b>*UCLA <math>\geq 5</math></b>   | 29.78%        | 94.06%        |
| $t^{(pos\%)} = 5$ (T2)             | 83.85%        | 42.86%        |
| $t^{(pos\%)} = NaN$ (T2)           | <b>80%</b>    | <b>41.98%</b> |
| $t^{(pos\%)} = 0.9$ (T2)           | 78.88%        | 41.73%        |
| $t^{(pos\%)} = 0.5$ (bpMR)         | 74.58%        | <b>55.56%</b> |
| $t^{(pos\%)} = NaN$ (bpMR)         | <b>80%</b>    | <b>44.13%</b> |
| $t^{(pos\%)} = 0.1$ (bpMR)         | 81.3%         | 44.44%        |

(c) Patient-level classification was performed using various UCLA scores as thresholds, both before and after AI assistance. The low specificity observed among radiologists (notated as \*) is anticipated due to the availability of labeled UCLA data. The ground truth of the UCLA data is biased towards positive patients.

Table 1: Different levels of classification. NaN indicates the values are computed though linear interpolation.  $^+ROIs$  indicates the radiologist-positive cases or lesions

## REFERENCES

- [1] Schelb, P., Kohl, S., Radtke, J. P., Wiesenfarth, M., Kickingeder, P., Bickelhaupt, S., Kuder, T. A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., et al., "Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment," *Radiology* **293**(3), 607–617 (2019).
- [2] Liu, S., Zheng, H., Feng, Y., and Li, W., "Prostate cancer diagnosis using deep learning with 3d multiparametric mri," in *[Medical imaging 2017: computer-aided diagnosis]*, **10134**, 581–584, SPIE (2017).
- [3] Dikaos, N., Alkalbani, J., Sidhu, H., Fujiwara, T., Abd-Alazeez, M., Kirkham, A., Allen, C., Ahmed, H. U., Emberton, M., Freeman, A., Halligan, S., Taylor, S., Atkinson, D., and Punwani, S., "Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric mri," *European radiology* **25** (09 2014).
- [4] Valerio, M., Anele, C., Freeman, A., Jameson, C., Singh, P., Hu, Y., Emberton, M., and Ahmed, H., "Identifying the index lesion with template prostate mapping biopsies," *The Journal of urology* **193** (11 2014).
- [5] Ahmed, H. U., El-Shater Bosaily, A., Brown, L. C., Gabe, R., Kaplan, R., Parmar, M. K., Collaco-Moraes, Y., Ward, K., Hindley, R. G., Freeman, A., Kirkham, A. P., Oldroyd, R., Parker, C., and Emberton, M.,

“Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study,” *The Lancet* **389**(10071), 815–822 (2017).

- [6] Natarajan, S., Priester, A., Margolis, D., Huang, J., and Marks, L., “Prostate mri and ultrasound with pathology and coordinates of tracked biopsy (prostate-mri-us-biopsy) (version 2).” Data set (2020). The Cancer Imaging Archive.
- [7] Zeevi, T., Leapman, M. S., Sprenkle, P. C., Venkataraman, R., Staib, L. H., and Onofrey, J. A., “Reliable prostate cancer risk mapping from mri using targeted and systematic core needle biopsy histopathology,” *IEEE Transactions on Biomedical Engineering* (2023).
- [8] He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., and Xu, D., “Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation,” in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*], Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., and Taylor, R., eds., 416–426, Springer Nature Switzerland, Cham (2023).