Decoupled conditional contrastive learning with variable metadata for prostate lesion detection

Camille Ruppli, Pietro Gori, Roberto Ardon, Isabelle Bloch

Decoupled conditional contrastive learning with variable metadata for prostate lesion detection

Camille Ruppli[1,3], Pietro Gori[1], Roberto Ardon[3], and Isabelle Bloch[2,1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France
[2] Sorbonne Université, CNRS, LIP6, Paris, France
[3] Incepto Medical, Paris, France

Abstract. Early diagnosis of prostate cancer is crucial for efficient treatment. Multi-parametric Magnetic Resonance Images (mp-MRI) are widely used for lesion detection. The Prostate Imaging Reporting and Data System (PI-RADS) has standardized interpretation of prostate MRI by defining a score for lesion malignancy. PI-RADS data is readily available from radiology reports but is subject to high inter-reports variability. We propose a new contrastive loss function that leverages weak metadata with multiple annotators per sample and takes advantage of inter-reports variability by defining metadata confidence. By combining metadata of varying confidence with unannotated data into a single conditional contrastive loss function, we report a 3% AUC increase on lesion detection on the public PI-CAI challenge dataset.
Code is available at : https://github.com/camilleruppli/decoupled_ccl

Keywords: Contrastive Learning · Semi-supervised Learning · Prostate cancer segmentation

1 Introduction

Clinical context. Prostate cancer is the second most common cancer in men worldwide. Its early detection is crucial for efficient treatment. Multi-parametric MRI has proved successful to increase diagnosis accuracy [21]. Recently, deep learning methods have been developed to automate prostate cancer detection [3,22,33]. Most of these methods rely on datasets of thousands of images, where lesions are usually manually annotated and classified by experts. This classification is based on the Prostate Imaging Reporting and Data System (PI-RADS) score, which ranges between 1 and 5, and associates a malignancy level to each lesion or to the whole exam (by considering the highest lesion score) [27]. This score is widely used by clinicians and is readily available from radiology reports. However, it is rather qualitative and subject to low inter-reader reproducibility [25]. Images can also be classified using biopsy results, as in the PI-CAI dataset [23]. This kind of classification is usually considered more precise (hence often taken as ground truth), but is also more costly to obtain and presents a bias since only patients with high PI-RADS scores undergo a

biopsy. Building a generic and automatic lesion detection method must therefore deal with the diversity of classification sources, radiology or biopsy, and the variability of classifications for a given exam.

Methodological context. In the past years, the amount of available medical imaging data has drastically increased. However, images are often either unannotated or weakly-annotated (*e.g.,* a single PI-RADS score for the entire exam), as annotating each lesion is costly and time consuming. This means that usual supervised models cannot be used, since their performance highly depends on the amount of annotated data, as shown in Table 1.

Table 1: AUC at exam level (metrics defined in Section 3) on a hold out test set of the PI-CAI dataset of models trained from random initialization with 5-fold cross validation on a private dataset.

	100% (N_{train}=1397)	10% (N_{train}=139)	1% (N_{train}=13)
3D UNet	0.80 (0.03)	0.76 (0.02)	0.71 (0.04)
3D ResUnet	0.79 (0.01)	0.73 (0.02)	0.64 (0.03)

To take advantage of unannotated or weakly-annotated data during a pretraining step, self-supervised contrastive learning methods [7,15,16] have been developed. Recent works have proposed to condition contrastive learning with class labels [18] or weak metadata [8,10,26] to improve latent representations. Lately, some works have also studied the robustness of supervised contrastive learning against noisy labels [30] proposing a new regularization [32].

While contrastive pretraining has been widely applied to classification problems [7,13,15,16], there have been few works about segmentation [2,6]. Recent works [1,6,35] propose to include pseudo labels at the pixel (decoder) level and not after the encoder, but, due to high computational burden, they can only consider 2D images and not whole 3D volumes.

Furthermore, many datasets contain several weak metadata at the exam level (e.g., PI-RADS score) obtained by different annotators. These weak metadata may have high inter- and intra-annotator variability, as for the PI-RADS score [25]. This variability is rarely taken into account into self-supervised pretraining. Researchers usually either use all annotations, thus using the same sample several times, or they only use confident samples, based on the number of annotators and their experience or on the learned representations, as in [19].

Contributions. Here, we aim to train a model that takes as input multiparametric MRI exam and outputs a map where higher values account for higher lesion probability. *Annotations* are provided by multiple annotators in the form of binary maps: segmentations of observed lesions. Only a small portion of the dataset has annotations. A greater proportion has *metadata* information available from written reports. These metadata, referring to the whole exam, are either a (binary) biopsy grading (presence or absence of malignant lesion) or a

PI-RADS score. For each exam, reports with a PI-RADS score are available from several radiologists, but the number of radiologists may differ among exams.

In the spirit of [5,11], we propose to include confidence, measured as a degree of inter-reports variability on metadata, in a contrastive learning framework. Our contributions are the following:

- We propose a new contrastive loss function that leverages weak metadata with multiple annotators per sample and takes advantage of inter-annotators variability by defining metadata confidence.
- We show that our method performs better than training from random initialization and previous pre-training methods on both the PI-CAI [24] public dataset and on a private multi-parametric prostate MRI dataset for prostate cancer lesion detection.

2 Method

We propose to apply contrastive pretraining to prostate lesion detection. A lesion is considered detected if the overlap between the predicted lesion segmentation and the reference segmentation is above 0.1, as defined in [22]. The predicted lesion masks are generated by a U-Net model [20] (since in our experiments this was the best model, see Table 1) fine-tuned after contrastive pretraining. In this section, we describe our contrastive learning framework defining confidence on metadata.

2.1 Contrastive learning framework

Contrastive learning (CL) methods train an encoder to bring close together latent representations of images of a positive pair while pushing further apart those of negative pairs. In unsupervised CL [7], where no annotations or metadata are available, a positive pair is usually defined as two transformations of the same image and negative pairs as transformed versions of different images. Transformations are usually randomly chosen among a predefined family of transformations. The final estimated latent space is structured to learn invariances with respect to the applied transformations.

In most CL methods, latent representations of different images are pushed apart *uniformly*. The alignment/uniformity contrastive loss function proposed in [28] is:

$$\mathcal{L}_{NCE} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} d_{ii}}_{\text{Global Alignment}} + \underbrace{\log\left(\frac{1}{N^2}\sum_{i,j=1}^{N} e^{-d_{ij}}\right)}_{\text{Global Uniformity}} \tag{1}$$

where $d_{ij} = ||x_1^i - x_2^j||_2$, x_1^i and x_2^j are the encoder outputs of the transformed versions of images i and j, respectively. However, as in many medical applications, our dataset contains discrete clinical features as metadata: PI-RADS scores and biopsy results per exam, that should be used to better define negative and

positive samples. To take metadata into account in contrastive pretraining, we follow the work of [9,10]. The authors introduce a kernel function on metadata y to *condition* positive and negative pairs selection, defining the following loss function:

$$\mathcal{L}_w = \underbrace{\frac{1}{N}\sum_{i,j=1}^{N} w(y_i, y_j) d_{ij}}_{\text{Conditional Alignment}} + \underbrace{\log\big(\frac{1}{N^2}\sum_{i,j=1}^{N}(||w||_\infty - w(y_i, y_j))e^{-d_{ij}}\big)}_{\text{Conditional Uniformity}} \tag{2}$$

where w is a kernel function measuring the degree of similarity between metadata y_i and y_j, $0 \leq w \leq 1$ and $||w||_\infty = w(y_i, y_i) = 1$. The conditional alignment term brings close together, in the representation space, only samples that have a metadata similarity greater than 0, while the conditional uniformity term does not repel all samples uniformly but weights the repulsion based on metadata dissimilarity. A schematic view of these two objective functions is shown in Figure 1 of the supplementary material.

We apply this framework to metadata (PI-RADS scores and biopsy results) that can have high inter-report variability.

To simplify the problem and homogenize PI-RADS and biopsy scores, we decide to binarize both scores, following clinical practice and medical knowledge [12,27]. We set $y = 0$ for PI-RADS 1 and 2, and $y = 1$ for PI-RADS 4 and 5. We do not consider PI-RADS 3, since it has the highest inter-reader variability [14] and low positive predictive value [29]. This means that all exams with a PI-RADS 3 are considered deprived of metadata. For each exam i, a set of y values is available, noted $\mathbf{y}_i$. The number of annotations may differ among subjects (see Equation (5) for a definition of w in such cases). For a biopsy result (defining an ISUP classification [12]), we set $y = 0$ if ISUP ≤ 1 and $y = 1$ if ISUP ≥ 2.

To take advantage of the entire dataset, we also consider unannotated data for which metadata are not provided. When computing the loss function on an exam without metadata (no $\mathbf{y}$ associated), we use the standard (unsupervised) contrastive loss function, as defined in [7]. This leads to the following contrastive loss function:

$$\mathcal{L}_w = \left.\begin{array}{l} \frac{1}{|A|}\sum_{i\in A}\Big(\sum_{j\in A} w(\mathbf{y}_j, \mathbf{y}_i)||x_1^i - x_2^j||\Big) \\ +\log\big(\frac{1}{|A|^2}\sum_{i,j\in A}(1 - w(\mathbf{y}_i, \mathbf{y}_j))e^{-||x_1^i - x_2^j||}\big) \end{array}\right\}_{\text{with metadata}} \\ \left.\begin{array}{l} +\frac{1}{|U|}\sum_{i\in U}\Big(||x_1^i - x_2^i||\Big) + \log\Big(\frac{1}{|U|^2}\sum_{\substack{i,j\in U \\ i\neq j}} e^{-||x_1^i - x_2^j||}\Big) \end{array}\right\}_{\text{without metadata}} \tag{3}$$

where A (resp. U) is the subset with (resp. without) associated $\mathbf{y}$ metadata.

Since the number of annotations may be different between two subjects i and j, we cannot use a standard kernel, as the RBF in [10]. We would like to take into account metadata confidence, namely agreement among annotators. In

the following, we propose a new kernel w that takes metadata confidence into account.

Confidence. Our measure of confidence is based on the discrepancy between the elements of vector $\mathbf{y}$ and their most common value (or majority voting). For exam i, if y_i is the most common value in its metadata vector $\mathbf{y}_i = [y_{i0}, y_{i1}, ...y_{in-1}]$ with n the number of available scores, confidence c is defined as:

$$c(\mathbf{y}_i) = \begin{cases} \epsilon & \text{if } n = 1 \\ 2 \times \left(\frac{\sum_{k=0}^{n-1} \delta(y_{ik}, y_i)}{n} - \frac{1}{2} \right) & \text{if } n > 1 \end{cases} \tag{4}$$

where δ is the Dirac function and $\epsilon = 0.1$[1]. $c(\mathbf{y}_i) \in [0, 1]$, 0 is found when an even number of opposite scores is obtained and the majority voting cannot provide a decision. In that case the associated exam will be considered as deprived of metadata. The proposed kernel then reads :

$$w(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} 1 & \text{if } i = j \quad \text{(exam against its own transformed version)} \\ c_{ij} & \text{if } y_i = y_j \text{ and } i \neq j \quad \text{(different exams, same majority voting)} \\ 0 & \text{if } y_i \neq y_j \text{ and } i \neq j \quad \text{(different exams, different majority voting)} \end{cases} \tag{5}$$

where $c_{ij} = \min(c(\mathbf{y}_i), c(\mathbf{y}_j))$. For two given exams i and j, the proposed model is interpreted as follows:

- If both metadata confidences are maximal ($c_{ij} = 1$), $w(\mathbf{y}_i, \mathbf{y}_j)$ will be equal to 1 and full alignment will be computed.
- If either metadata confidence is less than 1, $w(\mathbf{y}_i, \mathbf{y}_j)$ value will be smaller and exams will not be fully aligned in the latent space. The less confidence, the less aligned exams i and j representations will be.
- If confidence drops to zero for either exam, the exam will only be aligned with its own transformed version.

Similarly to decoupled CL [31], we design w such that the second term of Equation (3) does not repel samples with identical metadata most common value and maximal confidence ($c_{ij} = 1$). See Figure 1 for a schematic view.

[1] The maximum number of metadata available for an exam is $n = 7$, the minimal achievable confidence value is thus $c = 2(4/7 - 1/2) > 0.14$. We fix ϵ so that the confidence for $n = 1$ is higher than 0 but less that the minimal confidence when n is odd.

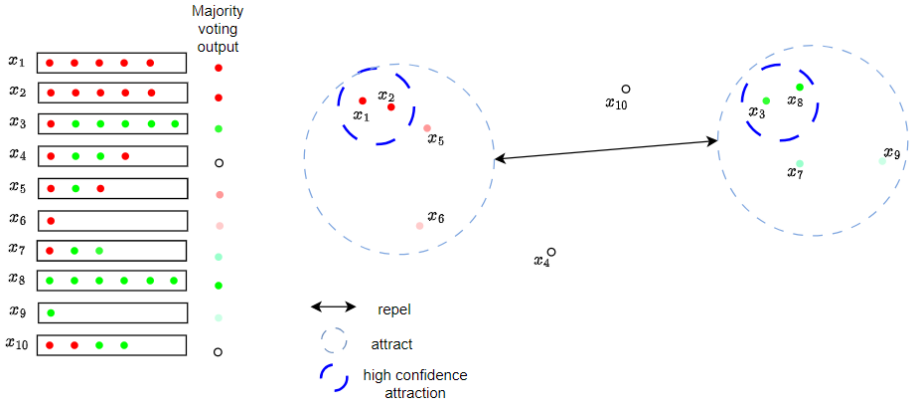

Fig. 1: Given a set of exams $x_{i\in[1,10]}$, $\mathbf{y}_i$ is represented as a list of colored points. Confidence (c) is represented with color saturation: darker means more confident. Exams such that $c(\mathbf{y}_i) = 0$ (no decision from majority voting) are considered as unlabeled and uncolored. Exams such that $c(\mathbf{y}_{i,j}) = 1$ and $y_i = y_j$, e.g. (x_1, x_2) (resp. (x_3, x_8)), will be strongly attracted while less attracted to patients with $c(\mathbf{y}_i) < 1$, e.g. $x_{5,6}$ (resp. $x_{7,9}$). Groups of exams with different y scores are repelled.

2.2 Experimental settings

Datasets. Experiments were performed on a private dataset of 2415 multi parametric MRI prostate exams among which 1397 have annotations (metadata and manual lesion segmentation) provided by multiple radiologists (up to 7). We also use the public PI-CAI dataset [23] composed of 1500 exams and 1295 annotations. In all learning steps, we used T2 weighted (T2w), apparent diffusion coefficient (ADC) and diffusion weighted (with the highest available b-value in the exam) sequences. As in [22,34], we use the prostate ternary segmentation (background, peripheral zone and central zone), generated from an independent process on T2w sequences. We thus learn from a total of four volumes considered as registered. Pretraining is performed on data from both datasets on 3915 exams. Fine-tuning is performed with 1% and 10% of these exams using cross validation (see **Implementation details**).

Implementation details. We pretrain the chosen U-Net encoder followed by a projection head. Similarly to nn U-Net [17], the encoder is a fully convolutional network where spatial anisotropy is used (e.g. axial axis is downsampled with a lower frequency since MRI volumes often have lower resolution in this direction). It is composed of four convolution blocks with one convolution layer in each block and takes the four sequences as input in channel dimension. The projection head is a two-layer perceptron as in [7]. We train with a batch size of 16 for 100 epochs and use a learning rate of 10^{-4}. Following the work of [13] on contrastive learning

for prostate cancer triage, we use a random sampling of rotation, translation and horizontal flip to generate the transformed versions of the images.

To evaluate the impact of contrastive pretraining at low data regime, we perform fine-tuning with 10% and 1% of annotated exams. The contrastive pretrained encoder is used to initialize the U-Net encoder, the whole encoder-decoder architecture is then fine-tuned on the supervised task. Fine-tuning is performed with **5-fold cross validation** with both datasets using the pretrained encoder. Using 1% (resp. 10%) of annotated data, each fold has 39 (resp. 269) training data and 12 (resp. 83) validation data. We build a hold out test set of 500 volumes[2], not used during any training step with data from both datasets to report our results. We also compared fine-tuning from our pretrained encoder to a model trained from random initialization. Fine-tuning results with 10% of annotated data are reported the in supplementary material.

Computing infrastructure. Optimizations were run on GPU NVIDIA T4 cards.

3 Results and discussion

The 3D U-Net network outputs lesions segmentation masks which are thresholded, following the dynamic thresholding proposed in [4], and of which connected components are computed. For each connected component, a detection probability is assigned as the maximum value of the network output in this component. The output of this post-processing is a binary mask associated with a detection probability per lesion. We compute the overlap between each lesion mask and the reference mask. A lesion is considered as a true positive (detection) if the overlap with the reference is above 0.1 as defined in [22]. This threshold is chosen to keep a maximum number of lesions to be analyzed for AUC computation. Different thresholds values are then applied for AUC computation.

As in [22,33], lesion detection probability is used to compute AUC values at exam and lesion levels, and average precision (mAP). To compute AUC at exam level we take, as ground truth, the absence or presence of a lesion mask, and, as a detection probability, the maximum probability of the set of detected lesions. At lesion level, all detection probabilities are considered and thresholded with different values, which amounts to limiting the number of predicted lesions. The higher this threshold, the lower are sensitivity and the number of predicted lesions, and the higher is specificity.

Results are presented in Table 2. For both datasets, we see that including metadata confidence to condition alignment and uniformity in contrastive pretraining yields better performances than previous state of the art approaches and random initialization. The discrepancy between PI-CAI and private mAP is due to the nature of the dataset: the PI-CAI challenge was designed to detect lesions confirmed by biopsy, while our private dataset contains lesions not necessarily confirmed by biopsy. Our private dataset contains manually segmented

[2] The 100 validation cases on the PI-CAI challenge website being hidden we could not compare our methods to the leaderboard performances.

lesions that might be discarded if biopsy was performed. The model being finetuned on both datasets, PI-CAI exams are overly segmented which leads to lower mAP values (since our model tends to over-segment on biopsy ground truths). For our clinical application, which aims to reproduce radiologist responses, this is acceptable. We report significant performance improvement at very low data regime (1% annotated data) compared to existing methods which is a framework often encountered in clinical practice.

To assess the impact of our approach we perform different ablation studies (shown in the second part of Table 2).

High confidence (HC row in Table 2). For pretraining, only exams with confidence equal to 1 are considered but are not perfectly aligned ($c_{ij} = 0.8\delta(c_{ij}, 1)$ in Equation (5)). We can see that considering only confident samples to condition contrastive learning decreased performances.

Majority Voting (Majority voting row in Table 2). We removed confidence and used majority voting output for kernel computation. If two different exams have the same majority vote we set : $w(y_i, y_j) = 0.8$ in Equation (5), other w values are kept unchanged. We can see that using majority voting output without taking confidence into account leads to decreased performances.

Biopsy (Biopsy row in Table 2). We set the confidence of PI-CAI exams to 1 (increasing biopsy confidence) which amounts to setting $\epsilon = 1$ for PI-CAI exams in Equation (4). No particular improvement is observed with this approach.

Global uniformity. We remove the conditioning on uniformity. Exams are uniformly repelled rather than conditioning on metadata similarity for repulsion (which amounts to setting $w(\mathbf{y}_i, \mathbf{y}_j) = 0$ for the second term of Equation (3)). Removing uniformity conditioning yields lower performances than the proposed approach (GlU row in Table 2).

Figure 2 shows the impact of our pretraining method on the finetuned U-Net outputs. Without conditioning, some lesions are missed (cases FN 1, FN 2) and others are falsely detected (cases FP 1, 2 and 3). Adding the conditioned pretraining removed these errors. More examples are provided in the supplementary material.

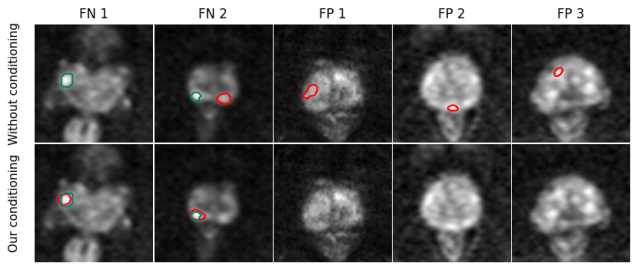

Fig. 2: Examples of false negative (FN) and false positive (FP) cases (first row) corrected by the proposed method (second row). Reference segmentation: green overlay, predicted lesions: red overlay

Table 2: 5-fold cross validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 1% of annotated data (standard deviation in parentheses)

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.68 (0.06)	0.74 (0.03)	0.73 (0.11)	0.70 (0.05)	0.27 (0.05)	0.62 (0.03)
Unif Align [28]	0.66 (0.07)	0.72 (0.01)	0.64 (0.13)	0.68 (0.03)	0.28 (0.07)	0.63 (0.03)
simCLR [7]	0.64 (0.07)	0.73 (0.05)	0.65 (0.08)	0.68 (0.05)	0.22 (0.07)	0.60 (0.06)
MoCo [16]	0.63 (0.08)	0.71 (0.04)	0.59 (0.12)	0.64 (0.07)	0.24 (0.10)	0.58 (0.06)
BYOL [15]	0.67 (0.06)	0.72 (0.04)	0.66 (0.16)	0.68 (0.04)	0.26 (0.05)	0.59 (0.04)
nnCLR [11]	0.57 (0.08)	0.73 (0.05)	0.49 (0.09)	0.62 (0.06)	0.21 (0.05)	0.59 (0.05)
Ours	**0.70** (0.05)	**0.75** (0.03)	**0.75** (0.10)	0.71 (0.03)	**0.30** (0.09)	**0.63** (0.04)
GlU	0.60 (0.05)	0.74 (0.03)	0.60 (0.12)	**0.73** (0.02)	0.23 (0.05)	0.63 (0.04)
Biopsy	0.64 (0.06)	0.73 (0.03)	0.69 (0.06)	0.70 (0.03)	0.24 (0.05)	0.62 (0.04)
HC	0.66 (0.08)	0.75 (0.04)	0.60 (0.06)	0.67 (0.03)	0.28 (0.09)	0.62 (0.03)
Majority voting	0.63 (0.06)	0.74 (0.02)	0.62 (0.06)	0.69 (0.04)	0.28 (0.07)	0.61 (0.04)

4 Conclusion

We presented a new method to take the confidence of metadata, namely the agreement among annotators, into account in a contrastive pretraining. We proposed a definition of metadata confidence and a new kernel to condition positive and negative sampling. The proposed method yielded better results for prostate lesion detection than existing contrastive learning approaches on two datasets.

References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. ICCV pp. 8199–8208 (2021)
2. Basak, H., Yin, Z.: Pseudo-Label Guided Contrastive Learning for Semi-Supervised Medical Image Segmentation. CVPR (2023)
3. Bhattacharya, I., Seetharaman, A., et al.: Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework. Medical Image Analysis **75** (2021)
4. Bosma, J.S., Saha, A., et al.: Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI - arxiv:2112.05151 (2021)
5. Bovsnjak, M., Richemond, P.H., et al.: SemPPL: Predicting pseudo-labels for better contrastive representations. ICLR (2023)
6. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis **87**, 102792 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. vol. 119, pp. 1597–1607 (2020)

8. Dufumier, B., Barbano, C.A., Louiset, R., Duchesnay, E., Gori, P.: Integrating Prior Knowledge in Contrastive Learning with Kernel. In: International Conference on Machine Learning (ICML) (2023)
9. Dufumier, B., Gori, P., et al.: Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels. In: MedNeurIPS (2021)
10. Dufumier, B., Gori, P., et al.: Contrastive Learning with Continuous Proxy Metadata for 3D MRI Classification. In: MICCAI, vol. 12902, pp. 58–68 (2021)
11. Dwibedi, D., Aytar, Y., et al.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. ICCV pp. 9568–9577 (2021)
12. Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A.: The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. The American Journal of Surgical Pathology **40**, 244–252 (2015)
13. Fernandez-Quilez, A., Eftestøl, T., Kjosavik, S.R., Olsen, M.G., Oppedal, K.: Contrasting axial T2W MRI for prostate cancer triage: A self-supervised learning approach. ISBI pp. 1–5 (2022)
14. Greer, M.D., et al.: Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. AJR pp. 1–8 (2019)
15. Grill, J.B., Strub, F., et al.: Bootstrap your own latent: A new approach to self-supervised learning. NeurIPS (2020)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. CVPR pp. 9726–9735 (2019)
17. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203 – 211 (2020)
18. Khosla, P., Teterwak, P., et al.: Supervised contrastive learning. NeurIPS (2020)
19. Li, S., Xia, X., Ge, S., Liu, T.: Selective-supervised contrastive learning with noisy labels. In: CVPR. pp. 316–325 (2022)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
21. Rouvière, O., et al.: Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. The Lancet. Oncology **20 1**, 100–109 (2019)
22. Saha, A., Hosseinzadeh, M., Huisman, H.J.: End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effect of attention mechanisms, clinical priori and decoupled false positive reduction. Medical Image Analysis **73**, 102155 (2021)
23. Saha, A., Twilt, J.J., et al.: Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (2022)
24. Saha, A., Twilt, J.J., et al.: The PI-CAI Challenge: Public Training and Development Dataset (May 2022)
25. Smith, C.P., et al.: Intra- and interreader reproducibility of PI-RADSv2: A multireader study. Journal of Magnetic Resonance Imaging **49** (2019)
26. Tsai, Y.H.H., Li, T., et al.: Conditional contrastive learning with kernel. ICLR (2022)
27. Turkbey, B.I., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European Urology (2019)

28. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML. vol. 119, pp. 9929–9939 (2020)
29. Westphalen, A.C., et al.: Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the society of abdominal radiology prostate cancer disease-focused panel. Radiology p. 190646 (2020)
30. Xue, Y., Whitecross, K., Mirzasoleiman, B.: Investigating why contrastive learning benefits robustness against label noise. In: ICML. pp. 24851–24871 (2022)
31. Yeh, C., Hong, C., et al.: Decoupled contrastive learning. In: ECCV. pp. 668–684 (2022)
32. Yi, L., Liu, S., She, Q., McLeod, A., Wang, B.: On learning contrastive representations for learning with noisy labels. CVPR pp. 16661–16670 (2022)
33. Yu, X., Lou, B., et al.: Deep attentive panoptic model for prostate cancer detection using biparametric MRI scans. In: MICCAI (2020)
34. Yu, X., Lou, B., et al.: False positive reduction using multiscale contextual features for prostate cancer detection in multi-parametric MRI scans. IEEE 17th International Symposium on Biomedical Imaging (ISBI) pp. 1355–1359 (2020)
35. Zhao, X., Vemulapalli, R., Mansfield, P.A., Gong, B., Green, B., Shapira, L., Wu, Y.: Contrastive learning for label efficient semantic segmentation. ICCV pp. 10603–10613 (2020)