Joint multi-task learning improves weakly-supervised biomarker prediction in computational pathology

Omar S. M. El Nahhas[1,2], Georg Wölflein[3], Marta Ligero[1], Tim Lenz[1], Marko van Treeck[1], Firas Khader[2,4], Daniel Truhn[2,4], and Jakob Nikolas Kather[1,2,5,6]

[1] Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, TUD Dresden University of Technology, Germany
[2] StratifAI GmbH, Dresden, Germany
[3] School of Computer Science, University of St Andrews, St Andrews, United Kingdom
[4] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen Germany
[5] Department of Medicine 1, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Germany
[6] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Abstract. Deep Learning (DL) can predict biomarkers directly from digitized cancer histology in a weakly-supervised setting. Recently, the prediction of continuous biomarkers through regression-based DL has seen an increasing interest. Nonetheless, clinical decision making often requires a categorical outcome. Consequently, we developed a weakly-supervised joint multi-task Transformer architecture which has been trained and evaluated on four public patient cohorts for the prediction of two key predictive biomarkers, microsatellite instability (MSI) and homologous recombination deficiency (HRD), trained with auxiliary regression tasks related to the tumor microenvironment. Moreover, we perform a comprehensive benchmark of 16 task balancing approaches for weakly-supervised joint multi-task learning in computational pathology. Using our novel approach, we outperform the state of the art by +7.7% and +4.1% as measured by the area under the receiver operating characteristic, and enhance clustering of latent embeddings by +8% and +5%, for the prediction of MSI and HRD in external cohorts, respectively.

Keywords: Pathology · Joint-learning · Multi-task · Weakly-supervised

1 Introduction

Over the past years, Deep Learning (DL) has proven its utility in predicting biomarkers directly from WSIs with hematoxylin- and eosin (H&E)-staining in a weakly-supervised manner. Weakly-supervised learning in computational

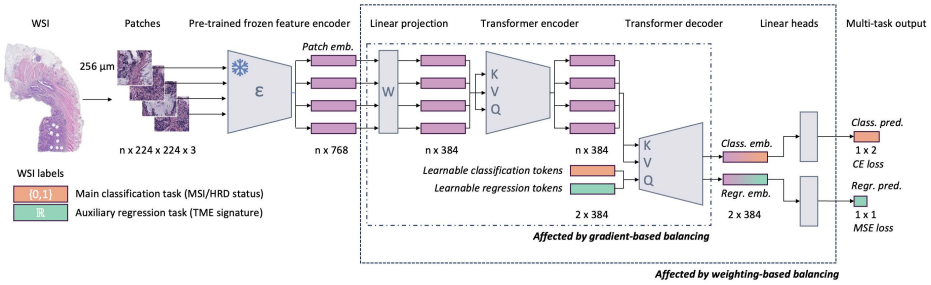

Fig. 1. Model overview. We tessellate WSIs into patches, extract CTransPath features [25], linearly project them, and feed them into a Transformer encoder. A learnable classification and regression token are added to the input of the Transformer decoder, after which the output is fed to a classification and regression head, performing weakly-supervised joint multi-task learning with weighting- and gradient-based task balancing.

pathology allows for large-scale analyses using solely the reported diagnosis as training labels, eliminating the need for cost- and time expensive pixel-level annotations [2]. The majority of studies predict categorical biomarkers with classification-based methods [9, 24], with a recent study showing the benefit of applying a regression-based method instead of dichotomizing the target for reformulation as a classification problem [6]. The studies predominantly follow the same pattern for model validation, often using heatmaps, top tiles and concordance analyses to confirm the model's alignment with known biological concepts [19, 14]. For example, biomarkers such as microsatellite instability (MSI) and homologous recombination deficiency (HRD) are predictive biomarkers which have known correlations with immune cells in the tumor microenvironment (TME) [1, 20]. However, the current state of the art for predicting MSI and HRD do not use observations from the tumor microenvironment as an additional learned task [24, 6], potentially leaving room for improved biomarker prediction. This leads to our primary research question: *Does including additional biological information in the form of an auxiliary regression task improve the prediction performance of the main classification task in weakly-supervised computational pathology?* Consequently, we develop a joint multi-task learning Transformer model which predicts the main classification task of MSI or HRD, while learning additional information about the TME through an auxiliary regression task in a weakly-supervised setting. Our contributions are as follows:

1. We propose a weakly-supervised joint multi-task learning framework that allows for additional biological information about the tumor microenvironment to be learned to improve the main biomarker prediction objective.
2. We conduct the first comprehensive benchmark of 16 multi-task balancing approaches in weakly-supervised computational pathology.
3. We improve over state-of-the-art weakly-supervised classification models for 2 highly relevant biomarkers, MSI and HRD, in 4 publicly available cohorts. Furthermore, we publicly release our code to promote reproducibility.

2 Related work

The concept of multi-task learning has been applied to the field of computational pathology for H&E WSIs in various studies. Yan et al. [27] and Graham et al. [8] combined segmentation and classification tasks using cross-entropy (CE) losses which are summed with equal weights for each task. A variety of studies combined solely classification objectives in a multi-task setting, using CE losses which are equally summed across the tasks [21, 18], or as a weighted sum with constants found through a hyperparameter search [16, 17]. Gao et al. [7] combined a CE loss with a mean-squared error (MSE) loss which are balanced according to preset constants which only update in specific, pre-defined scenarios, and are manually bounded. Only Lu et al. [16] and Marini et al. [17] approached the multi-task problem from a weakly-supervised perspective. In summary, prior studies opted for weighted-based balancing approaches for multi-task learning, which were either equally balanced, or fine-tuned for very specific use-cases that likely do not translate well to other scenarios of a similar kind [21]. This leaves a clear gap in the computational pathology literature for the application of more sophisticated, model-guided balancing of losses and gradients [3, 10–13, 15, 28], especially in a weakly-supervised setting.

3 Method

We consider a dataset of N WSIs $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$, where each WSI $\mathbf{X}^{(i)} \in \mathbb{R}^{W \times H \times 3}$ is an RGB image of width W and height H, though these dimensions may vary between slides. During training, each WSI $\mathbf{X}^{(i)}$ is associated with a binary classification label $y^{(i)} \in \mathcal{Y} = \{0, 1\}$ for the main task, as well as an auxiliary regression label $a^{(i)} \in \mathbb{R}$. For example, the classification label $y^{(i)}$ could indicate MSI status, and the auxiliary target $a^{(i)}$ could represent a molecular signature for lymphocyte infiltration, which takes on continuous values.

Due to their large size, it is common to consider WSIs as collections of patches, framing the WSI classification problem as a weakly supervised learning task. More specifically, we split each WSI $\mathbf{X}^{(i)}$ into a set of n non-overlapping patches $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_n^{(i)}\}$ where each $\mathbf{x}_j^{(i)} \in \mathcal{X} = \mathbb{R}^{P \times P \times 3}$ for a fixed patch size P (the number of patches n varies depending on the particular slide's dimensions). We follow the STAMP protocol [5], which sets the patch size $P = 224$ at an edge length of 256 microns (which corresponds to approximately 9× magnification), yielding $n^{(i)} \in \mathbb{N}$ non-background patches per slide. The task is to train a model $M : \mathcal{P}(\mathcal{X}) \to \mathcal{Y}$ that at inference time predicts the classification label given a bag of patches representing a WSI. During training, this model should learn from both the classification labels y and the auxiliary regression target a, though at inference time we are only interested in the former.

Obtaining the prediction from a collection of patches representing a WSI is a two-step process consisting of (i) feature extraction and (ii) feature aggregation, outlined in Fig. 1. We describe these steps in the sections below. The source code is available at: https://github.com/KatherLab/joint-mtl-cpath.

3.1 Feature extraction

Our model operates on feature vectors instead of raw patches. Thus, we first apply a feature extractor $\mathcal{E} : \mathcal{X} \rightarrow \mathbb{R}^{d_z}$ individually to each patch $\mathbf{x}_j^{(i)}$ in order to obtain a corresponding feature vector $\mathbf{z}_j^{(i)} = \mathcal{E}\left(\mathbf{x}_j^{(i)}\right)$ that meaningfully represents each patch. We parameterize $\mathcal{E}$ with CTransPath [25], a model that was pretrained on 32,000 WSIs across various cancer types using self-supervised learning. The extracted CTransPath feature vectors, which are of dimensionality $d_z = 768$, are cached before training to save compute. As such, our preprocessing and feature extraction setup closely follows the STAMP [5] protocol, except for the exclusion of stain normalisation [26].

3.2 Architecture

Our joint multi-task Transformer architecture (Fig. 1) modifies Vaswani et al.'s design [23]. We first project features into a lower-dimensional latent space to manage complexity, then encode these projected tokens using a Transformer encoder stack. We decode using `[cls]` tokens for classification and `[rgr]` tokens for regression. Each decoded token passes through a fully connected layer for label-wise prediction. This architecture improves upon the classic Vision Transformer[4] for scalable multi-task, multi-label predictions. Key differences from Vaswani et al.[23] include: 1) an initial projection stage enabling use with larger input feature dimensions, and 2) fixed, learned class tokens with independent fully connected layers for multiple label prediction.

3.3 Training

All models are trained in a weakly-supervised setting using CTransPath feature vectors[25]. We employ 5-fold cross-validation with an 80-20 split for training and testing, maintaining consistent patient splits across all compared models. We report the mean AUROC and AUPRC with 95% confidence interval, and mean silhouette score (SS) across the 5 folds. The baseline model performs only classification of MSI or HRD, while the joint-learned model additionally regresses tumor microenvironment signatures: lymphocyte infiltrating signature score (LISS), leukocyte fraction (LF), stromal fraction (SF), tumor cell proliferation (Prolif), and intratumor heterogeneity (ITH), chosen for their biological concordance with MSI and HRD (Suppl. Fig. 1). We optimize using AdamW [15] with a learning rate of 1e-4, CE loss for classification, and MSE loss for regression. Training uses all n patch features per WSI with a batch size of 1 for 32 epochs. Early stopping is triggered after 7 epochs without CE loss decrease in the primary classification task.

3.4 Multi-task balancing

We apply and compare a total of 16 task balancing approaches for the joint multi-task learning experiments. For weighting-based balancing, we use uncertainty (*uncert*) [10], dynamic weight averaging (*dwa*) [13], and Auto-Lambda

(*autol*) [12]. For gradient-based balancing, we use gradient sign dropout (*graddrop*) [3], projecting conflicting gradients (*pcgrad*) [28], and conflict-averse gradient descent (*cagrad*) [11]. For comparison with methods used in prior studies, we include an approach which weights the tasks equally (*naive*). Previous work states that combining weighting- and gradient-based balancing in multi-task learning can improve performance [12], which leads to the combination of aforementioned methods. All balancing approaches focus on single objective optimization, i.e. improving the classification performance regardless of the regression performance, except for *autol* which performs multi-objective optimization for both classification and regression [12]. Weighting-based balancing affects all non-frozen layers in the network, whereas gradient-based balancing only affects the shared projector, encoder and decoder layers (Fig. 1).

4 Experiments and Results

4.1 Data

We use four public cohorts for the training and evaluation of the models. For MSI, we train on the colorectal cancer (CRC) cohort from The Cancer Genome Atlas (TCGA), TCGA-CRC, and evaluate on the CRC cohort from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), CPTAC-CRC. For HRD, we train on the lung adenocarcinoma (LUAD) cohort from TCGA, TCGA-LUAD, and evaluate on the LUAD cohort from CPTAC, CPTAC-LUAD. The public biomarker data for MSI is from the study by Wagner et al. [24], for HRD is from the study by El Nahhas et al. [6], and for the TME is from the study by Thorsson et al. [22].

4.2 Joint multi-task learning improves classification predictions

We develop a Transformer architecture for weakly-supervised joint multi-task classification and regression using WSI features. To our knowledge, this is the first work predicting MSI or HRD directly from WSIs in a joint multi-task setting. We compare our model to state-of-the-art MSI [24] and HRD [6] weakly-supervised classification models and include a baseline without auxiliary regression tasks. Our baseline outperforms the state-of-the-art MSI model (AUROC 86.1% vs 83.0% [24] in TCGA-CRC) and HRD model (AUROC 71.6% vs 70.0% [6] in TCGA-LUAD). Introducing auxiliary regression tasks for tumor microenvironment quantification further improves performance. Our joint multi-task model achieves an AUROC of 94.0% and AUPRC of 84.5% for MSI prediction in TCGA-CRC, and an AUROC of 73.4% and AUPRC of 59.8% for HRD prediction in TCGA-LUAD (Table 1, Suppl. Table 1 and 2). This represents improvements of +11% and +3.4% in AUROC over the state of the art, respectively. These results demonstrate that weakly-supervised joint multi-task learning enhances classification performance compared to both the baseline and state-of-the-art.

Table 1. Performance overview of weakly-supervised MSI and HRD biomarker prediction models comparing weighting- and gradient-based task balancing methods.

	MSI				HRD			
	TCGA-CRC		CPTAC-CRC		TCGA-LUAD		CPTAC-LUAD	
Methods	AUC	PRC	AUC	PRC	AUC	PRC	AUC	PRC
SOTA	83.0	-	82.0	-	70.0	-	82.0	-
baseline	86.1	61.4	86.4	70.5	71.6	57.7	81.0	30.3
naive	86.4	62.7	88.2	72.4	69.6	57.6	81.2	33.7
dwa	84.2	58.7	87.7	70.8	73.3	60.3	83.6	40.8
uncert	86.0	63.4	88.4	72.6	73.2	60.1	83.1	41.5
autol	**94.0**	**84.5**	86.9	73.1	72.2	58.5	85.2	43.0
graddrop	85.8	61.2	87.3	71.7	71.1	58.8	84.4	42.5
pcgrad	85.4	62.4	87.3	72.0	72.1	**60.4**	84.1	41.2
cagrad	86.5	62.7	89.7	**76.6**	72.6	58.7	85.4	42.1
dwa + graddrop	85.5	59.8	87.6	71.6	72.6	58.4	83.3	40.2
dwa + pcgrad	85.6	62.3	88.8	73.3	**73.4**	59.8	83.0	39.1
dwa + cagrad	85.8	59.9	88.4	73.8	71.8	57.9	85.5	44.3
uncert + graddrop	85.5	60.7	87.6	71.7	72.6	59.4	83.9	42.3
uncert + pcgrad	86.7	62.5	89.0	74.2	71.8	55.4	83.6	39.9
uncert + cagrad	86.3	60.7	88.9	74.6	72.4	58.8	84.4	41.9
autol + graddrop	85.3	61.6	86.7	69.4	70.8	56.0	84.8	43.4
autol + pcgrad	86.1	63.2	87.9	73.6	72.0	58.2	85.6	42.8
autol + cagrad	86.5	62.0	**89.9**	76.3	71.9	57.0	**86.1**	**43.8**

4.3 Joint multi-task learning improves generalizability

Next, we evaluate the generalizability of the joint multi-task learned models to external cohorts, comparing them to the baseline and state-of-the-art models for MSI [24] and HRD [6] classification (Table 1, Suppl. Table 1 and 2). The baseline model outperforms the state-of-the-art MSI model by +4.4% AUROC, while performing slightly worse (-1% AUROC) for HRD. Introducing auxiliary regression tasks with weighting- and gradient-balancing schemes substantially improves performance on external cohorts. The model with *autol + cagrad* balancing achieves an AUROC of 89.9% and AUPRC of 76.3% for MSI in CPTAC-CRC, and an AUROC of 86.1% and AUPRC of 43.8% for HRD in CPTAC-LUAD. This represents +7.7% and +4.1% improvements over state-of-the-art models for MSI and HRD prediction in external cohorts, respectively.

4.4 Joint multi-task learning improves latent-embedding clustering

Finally, we analyze the latent space of the classification head input (Fig. 1), comparing clustering capabilities of the *384*-dimensional embeddings in both classification and joint-learning settings. The joint-learned embeddings show superior clustering performance, with the best combinations being (*autol + cagrad*) for MSI (SS: 0.44) and (*dwa + cagrad*) or (*uncert + cagrad*) for HRD (SS: 0.12). These represent improvements of 8% and 5% over the baseline for MSI and HRD,

respectively. Visualization of embeddings using t-SNE in an external cohort (CPTAC-CRC) shows equal AUC (87%) but substantially improved SS (0.52 vs 0.33) for joint-learned embeddings compared to the baseline (Suppl. Fig. 2). These results demonstrate improved latent-embedding clustering and generalizability of our proposed model.

Table 2. Clustering performance of the latent embeddings on external cohorts as measured by the silhouette score.

	MSI CPTAC-CRC						HRD CPTAC-LUAD		
Methods	ITH	LF	LISS	Prolif	SF	mean	ITH	Prolif	mean
baseline	0.35	0.34	0.38	0.35	0.38	0.36	0.05	**0.10**	0.07
naive	0.43	0.39	0.42	0.41	**0.45**	0.42	0.05	0.08	0.08
dwa	0.26	0.33	0.33	0.33	0.27	0.30	0.07	0.01	0.04
uncert	0.30	0.34	0.39	0.33	0.31	0.33	0.07	0.01	0.04
autol	0.37	**0.48**	0.41	0.37	0.39	0.40	0.10	0.07	0.09
graddrop	0.31	0.35	0.29	0.32	0.27	0.31	0.08	0.04	0.06
pcgrad	0.28	0.44	0.38	0.32	0.35	0.35	0.08	0.01	0.05
cagrad	0.36	**0.48**	0.39	**0.45**	0.43	0.42	0.14	0.05	0.10
dwa + graddrop	0.31	0.31	0.33	0.31	0.27	0.31	0.04	0.03	0.04
dwa + pcgrad	0.31	0.39	0.38	0.38	0.30	0.35	0.05	0.03	0.04
dwa + cagrad	0.33	0.40	**0.45**	0.43	0.43	0.41	0.15	0.08	**0.12**
uncert + graddrop	0.31	0.31	0.32	0.31	0.23	0.30	0.08	0.03	0.06
uncert + pcgrad	0.33	0.44	0.34	0.38	0.33	0.36	0.05	0.02	0.04
uncert + cagrad	0.37	**0.48**	0.44	0.38	0.43	0.42	**0.16**	0.07	**0.12**
autol + graddrop	0.32	0.30	0.29	0.32	0.29	0.30	0.10	0.08	0.09
autol + pcgrad	0.34	0.37	0.31	0.35	0.32	0.34	0.10	0.06	0.08
autol + cagrad	**0.44**	0.45	0.43	**0.45**	0.41	**0.44**	0.12	0.07	0.10

5 Conclusion

We developed a weakly-supervised joint multi-task Transformer architecture incorporating tumor microenvironment information to improve MSI and HRD prediction. An ablation study of 16 multi-task balancing approaches demonstrated their impact on performance. Therefore, we conclude that a combination of gradient- and weighting-based balancing outperforms naive balancing and should be considered in multi-task weakly-supervised problems in computational pathology. Our approach achieves state-of-the-art performance in classifying MSI and HRD from WSIs across multiple cohorts, showing improved generalizability and latent space embedding clustering. This work demonstrates the potential of biology-informed deep learning with auxiliary tasks for predictive biomarker classification. Follow-up HRD studies should consider evaluation of our method on other solid tumors such as ovarian cancer.

Acknowledgments. This research has been partially funded by the German Federal Ministry of Education and Research (BMBF) through grant 1IS23070, Software Campus 3.0 (TU Dresden), as part of the Software Campus project 'MIRACLE-AI'.

Disclosure of Interests. OSMEN holds shares in StratifAI GmbH. FK holds shares in StratifAI GmbH. DT holds shares in StratifAI GmbH. JNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Scailyte, Switzerland; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore he holds shares in StratifAI GmbH, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. Remaining authors report no competing interests.

References

1. Bai, J., Chen, H., Bai, X.: Relationship between microsatellite status and immune microenvironment of colorectal cancer and its application to diagnosis and treatment. J. Clin. Lab. Anal. **35**(6), e23810 (Jun 2021)
2. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (Aug 2019)
3. Chen, Z., Ngiam, J., Huang, Y., Luong, T., Kretzschmar, H., Chai, Y., Anguelov, D.: Just pick a sign: Optimizing deep multitask models with gradient sign dropout (Oct 2020)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (Oct 2020)
5. El Nahhas, O.S.M., et al.: From whole-slide image to biomarker prediction: A protocol for End-to-End deep learning in computational pathology (Dec 2023)
6. El Nahhas, O.S.M., et al.: Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. Nat. Commun. **15**(1), 1–13 (Feb 2024)
7. Gao, Z., et al.: A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. Med. Image Anal. **83**, 102652 (Jan 2023)
8. Graham, S., et al.: One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Med. Image Anal. **83**, 102685 (Jan 2023)
9. Kather, J.N., et al.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. **25**(7), 1054–1056 (Jun 2019)
10. Kendall, A., Gal, Y., Cipolla, R.: Multi-Task learning using uncertainty to weigh losses for scene geometry and semantics (May 2017)
11. Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-Averse gradient descent for multi-task learning (Oct 2021)
12. Liu, S., James, S., Davison, A.J., Johns, E.: Auto-Lambda: Disentangling dynamic task relationships (Feb 2022)
13. Liu, S., Johns, E., Davison, A.J.: End-to-End Multi-Task learning with attention (Mar 2018)

14. Loeffler, C.M.L., et al.: Direct prediction of homologous recombination deficiency from routine histology in ten different tumor types with attention-based multiple instance learning: a development and validation study. medRxiv (Mar 2023)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (Sep 2018)
16. Lu, M.Y., et al.: AI-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (Jun 2021)
17. Marini, N., et al.: Multi-Scale task multiple instance learning for the classification of digital pathology images with global annotations. In: Proceedings of the MICCAI Workshop on Computational Pathology. Proceedings of Machine Learning Research, vol. 156, pp. 170–181. PMLR (Sep 2021)
18. Mormont, R., Geurts, P., Marée, R.: Multi-Task Pre-Training of deep neural networks for digital pathology. IEEE journal of biomedical and health informatics (2020)
19. Niehues, J.M., et al.: Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. Cell Rep Med p. 100980 (Mar 2023)
20. Shi, Z., Chen, B., Han, X., Gu, W., Liang, S., Wu, L.: Genomic and molecular landscape of homologous recombination deficiency across multiple cancer types. Sci. Rep. **13**(1), 8899 (Jun 2023)
21. Tellez, D., et al.: Extending unsupervised neural image compression with supervised multitask learning. In: Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 770–783. PMLR (2020)
22. Thorsson, V., et al.: The immune landscape of cancer. Immunity **48**(4), 812–830.e14 (Apr 2018)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
24. Wagner, S.J., et al.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell **41**(9), 1650–1661.e4 (Sep 2023)
25. Wang, X., et al.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **81**, 102559 (Oct 2022)
26. Wölflein, G., et al.: A good feature extractor is all you need for weakly supervised learning in histopathology (Nov 2023)
27. Yan, C., Xu, J., Xie, J., Cai, C., Lu, H.: Prior-Aware CNN with Multi-Task learning for colon images analysis. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 254–257. IEEE (Apr 2020)
28. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for Multi-Task learning (Jan 2020)