






# Patient-Level Microsatellite Stability Assessment from Whole Slide Images by Combining Momentum Contrast Learning and Group Patch Embeddings

Daniel Shats<sup>1</sup> , Hadar Hezi<sup>2</sup>, Guy Shani<sup>3</sup>, Yosef E. Maruvka<sup>3</sup> ,  
and Moti Freiman<sup>2</sup> 

<sup>1</sup> Faculty of Computer Science, Technion, Haifa, Israel  
shats@campus.technion.ac.il

<sup>2</sup> Faculty of Biomedical Engineering, Technion, Haifa, Israel

<sup>3</sup> Faculty of Biotechnology and Food Engineering, Technion, Haifa, Israel

**Abstract.** Assessing microsatellite stability status of a patient's colorectal cancer is crucial in personalizing treatment regime. Recently, convolutional-neural-networks (CNN) combined with transfer-learning approaches were proposed to circumvent traditional laboratory testing for determining microsatellite status from hematoxylin and eosin stained biopsy whole slide images (WSI). However, the high resolution of WSI practically prevent direct classification of the entire WSI. Current approaches bypass the WSI high resolution by first classifying small patches extracted from the WSI, and then aggregating patch-level classification logits to deduce the patient-level status. Such approaches limit the capacity to capture important information which resides at the high resolution WSI data. We introduce an effective approach to leverage WSI high resolution information by momentum contrastive learning of patch embeddings along with training a patient-level classifier on groups of those embeddings. Our approach achieves up to 7.4% better accuracy compared to the straightforward patch-level classification and patient level aggregation approach with a higher stability (AUC,  $0.91 \pm 0.01$  vs.  $0.85 \pm 0.04$ , p-value  $< 0.01$ ). Our code can be found at [https://github.com/TechnionComputationalMRILab/colorectal\\_cancer\\_ai](https://github.com/TechnionComputationalMRILab/colorectal_cancer_ai).

**Keywords:** Digital pathology · Colorectal cancer · Self-supervised learning · Momentum contrast learning

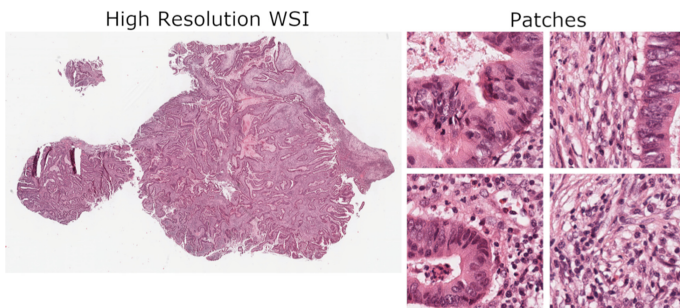
## 1 Introduction

Colorectal cancer is a heterogeneous type of cancer which can be generally classified into one of two groups based on the condition of Short Tandem Repeats (STRs) within the DNA of a patient. These two groups are known as Microsatellite Stable (MSS) and Microsatellite Instable (MSI). When the series

of nucleotides in the central cores of the STRs exhibit a highly variable number of core repeating units, the patient sample is referred to as Microsatellite Instable. Otherwise, if the number nucleotides is consistent, that patient sample is referred to as Microsatellite Stable (MSS) [18]. Microsatellite instability (MSI) exists in approximately 15% of colorectal cancer patients. Due to the fact that patients with MSI have different treatment prospects than those without MSI, it is critical to know if a patient exhibits this pathology before determining treatment direction [2].

Currently, it is possible to determine microsatellite status in a patient by performing various laboratory tests. While these methods are effective, they are expensive and take time, resulting in many patients not being tested for it at all. Therefore, there exist many recent efforts aimed toward utilizing deep-learning-based methods to uncover a computational solution for the detection of MSI/MSS status from hematoxylin and eosin (H&E) stained biopsy whole slide images (WSI).

However, the WSIs have an extremely large resolution, often reaching over a billion pixels (Fig. 1). Since neural networks can operate, due to computational constraints, only on relatively low resolution data, the input WSI must be shrunk down in some way to a size that is manageable by today’s models and hardware. A relatively naïve approach to do this is by down sampling inputs to approximately the same resolution as Image Net (which also allows one to leverage transfer learning) [10, 16]. While this is not detrimental for natural images, where fine detail might not be critical, the same cannot be said for WSI of human tissue where the information of concern may exist in various scales. Our work presents a novel method for effectively down sampling patches into lower dimensional embeddings while preserving features critical to making clinical decisions. Due to our effective dimensional reduction of individual patches, we are then able to concatenate the features of multiple patches together and make clinical decisions using inter-patch information.



**Fig. 1.** Example of an single patient WSI and a subset of the patches generated from the WSI. The resolution of the WSI is  $45815 \times 28384$  and the corresponding patches are  $512 \times 512$ .

### 1.1 CNN-Based MSI/MSS Classification from WSI Data

A simple and straightforward method to alleviate the above issue is by cutting WSI's into many patches and then applying modern deep learning methods to these individual patches as if each of them were a separate image from a data set [10,16]. Once the input has been tessellated, there are a few known ways of training a model from them. One way is to simply assign a label to every patch based on the patient from which they are derived, and then train a model on a patch-by-patch basis. Once the model has been trained, its outputs on all the patches can be averaged for a patient-level classification. Described mathematically, the inference procedure is described below:

Suppose we have some trained classifier  $F$  (which returns the probability that a patch belongs to class MSI or MSS e.g.  $0 \leq F(x) \leq 1$  for some input patch  $x$ ), a whole slide image  $W$ , and  $n$  patches extracted from  $W$  such that  $\{p_1, p_2, \dots, p_n\} \in W$ . Then, a patient level probability prediction  $P_W$  (on the WSI) can be formulated as such:

$$P_W = \frac{1}{n} \sum F(p_n) \quad (1)$$

And given some classification threshold  $0 \leq t \leq 1$ , we can arrive at a final classification  $C_W$  for the patient:

$$C_W = \begin{cases} MSS & P_W < t \\ MSI & P_W \geq t \end{cases} \quad (2)$$

However, such approaches practically ignore the fact that much of the information critical to making an informed decision on a patient level may reside in the high resolution and inter-patch space. Further, the classification of the patches based on the patient-level data may result in incorrect classification as not necessarily all patches are contributing equally to the classification of a patient as MSI or MSS.

### 1.2 Self-supervision for Patch Embeddings

In recent years self-supervised learning methods have become an extremely attractive replacement for autoencoders as encoding or downsampling mechanisms, while learning features that are much more informative and meaningful [19] and therefore can be easily leveraged for use in downstream tasks. This is an exciting property that has led to new research being done in an attempt to circumvent the issues with downsampling described earlier.

Although self-supervised learning is relatively new and there are many algorithms that attempt to achieve effectively the same goal [3,11,12]. Of the various methods, contrastive learning techniques such as those introduced in SimCLR [7] have become very popular due to their high efficacy and simplicity.

Recently there have been some attempts to use self-supervised contrastive learning, and more specifically the aforementioned SimCLR algorithm, to aid in

classification of WSI imagery. Unfortunately SimCLR requires large computational resources to train in a reasonable amount of time. That is why we decided to test the advent of Momentum Contrast Learning with MoCo v2 by Chen et al. [8]. This framework relies on storing a queue of previously encoded samples to overcome the large batch size requirement of SimCLR while seemingly improving downstream classification accuracy as well.

## 2 Related Work

### 2.1 CNN-Based MSS/MSI Classification

Building on top of Echle [10] and the Eqs. 1 and 2 was that of Kather et al. [16]. Here, transfer learning via pretraining with ImageNet [9] was employed to marginally improve results of this straightforward method. Although it has an enormous amount of images, they are not medical, and certainly not pictures of H&E stained biopsy slides. Therefore the degree to which the learned features from ImageNet transfer well to H&E stained WSI imagery is arguably negligible.

It is also important to discuss the particular resolution under which the patches were acquired. Due to the small size of the patches, any individual patch may not be large enough to contain the information required to make a classification (even on the tissue contained within only that patch). One must understand whether or not the task at hand requires intra-cellular information (requiring maximum slide resolution) or tissue-level information (requiring down-sampling before patching). Unfortunately in either case, it is also possible that information at multiple levels of resolution is required for optimal results.

Still, more drawbacks can be found tessellating high resolution images into many smaller patches, regardless of patch resolution concerns. For one, the model cannot learn inter-patch information. This is especially important considering that it is very possible that a majority of the patches do not actually contain targeted tissue. Moreover, training the model in such a way is misleading, considering that many patches which have MSS tissue (yet are found on an MSI classified patient) will be marked as MSI for the model. This is likely to result in the model learning less than optimal features.

The work by Hemati et al. [14] gets around this issue by creating a mosaic of multiple patches per batch during training. Unfortunately, this creates other drawbacks. Most notably, they cannot use all the patches per patient, and so they use another algorithm which is mutually exclusive to the learning procedure in order to choose patches, with no guarantee that they contain targeted tissue. Moreover, the mosaic of these selected patches is also still limited by resolution, and so they still must scale down the patches from their original resolution before training.

An improvement on all these previous works was done in the research by Bilal et al. [1]. Most notably, they advanced upon the work from Hemati [14] by learning the patch extraction, or as they call it, patch detection, using a neural network as well. Thereby alleviating an inductive bottleneck. Their process also includes significant work surrounding intermittent detection of known

biologically important features to such a problem, such as doing nuclear segmentation, and then providing that information to the next model to make a better-informed decision.

## 2.2 Self Supervision for Patch Embeddings

Due to self-supervised learning being a fairly recent invention, the works similar to ours which cite using it are rather sparse. One of the works which explores the validity of using these methods in the first place is that of Chen et al. [6]. They show that features learned through self-supervised contrastive learning are robust and can be learned from in a downstream fashion. Another paper that uses a similar two stage approach with self-supervised learning being the first stage is DeepSMILE by Schirris et al. [21]. They used a similar approach to the above mentioned contrastive self-supervised learning step with SimCLR, but learned on the features using Deep Multiple Instance Learning (MIL) [15]. While this approach was effective, the computational requirements of SimCLR and the added complexity of MIL may keep the advent of this research out of the hands of many researchers.

Very recently an improvement on the work by Chen [6] was introduced in their research using Hierarchical Vision transformers [5]. Here, the authors apply self-supervised learning through DINO [4] to train 3 stages of vision transformers in achieving entire WSI level classifications. Though seemingly effective, the increased complexity of their approach and the necessity of utilizing transformers makes it relatively inflexible.

## 2.3 Hypothesis and Contributions

*Hypothesis:* Learning effective patch embeddings with self-supervised learning and training a small classifier on groups of those embeddings is more effective than either training on down-sampled WSI's or training on individual patch embeddings and averaging the classification for a patient.

We believe this hypothesis to be true due to the ability of a network to learn inter-patch information at an embedding level. This way, information that is encoded in one patch can impact the decision of the entire WSI. Our contribution is an intuitive and elegant framework that improves patient classification accuracy by up to 7.4%. We argue that this method is very simple to understand and has many avenues of possible improvement. Specifically when considering the initial feature extraction stage, there are many other self-supervised representation learning methods that can be tested and directly compared using our approach.

## 3 Data

The training and validation data used in our method consists of the COAD and READ cohorts of The Cancer Genome Atlas (TCGA) [22]. Out of a total of 360

unique patients, the train set is comprised of 260 patients and the validation set is comprised of 100 patients, where each patient is equivalent to one WSI. Each of these WSIs were tessellated into patches of size  $512 \times 512$  pixels at  $0.5 \mu\text{m}/\text{px}$  and then downsampled to  $224 \times 224$  (this was done only for comparison with Kather et al. [16]). Next the probability of each patch to contain cancerous tissue was computed by a trained CNN and only the patches which were likely to have cancer tissue were kept. Finally, the patches were color normalized. Further detail on the data preprocessing procedure can be found in the paper by Kather et al. [17]. The 260 train patients were tessellated into 46,704 MSS labeled patches, and 46,704 MSIMUT labeled patches. The 100 validation patients were tessellated into 70,569 MSS labeled patches, and 28,335 MSIMUT labeled patches.

Finally, we also ran some final experiments on a more balanced subset of the validation dataset, referred to later in this paper as the ‘‘Balanced Validation Set’’. It comprises of 15 MSS patients and 15 MSIMUT patients which all have a relatively similar distribution of patches extracted from them. 7281 patches were extracted from the MSIMUT patients and 7446 patches were extracted from the MSS patients.

## 4 Method and Model

### 4.1 Overview

Our method comprises of two main training stages (Fig. 2):

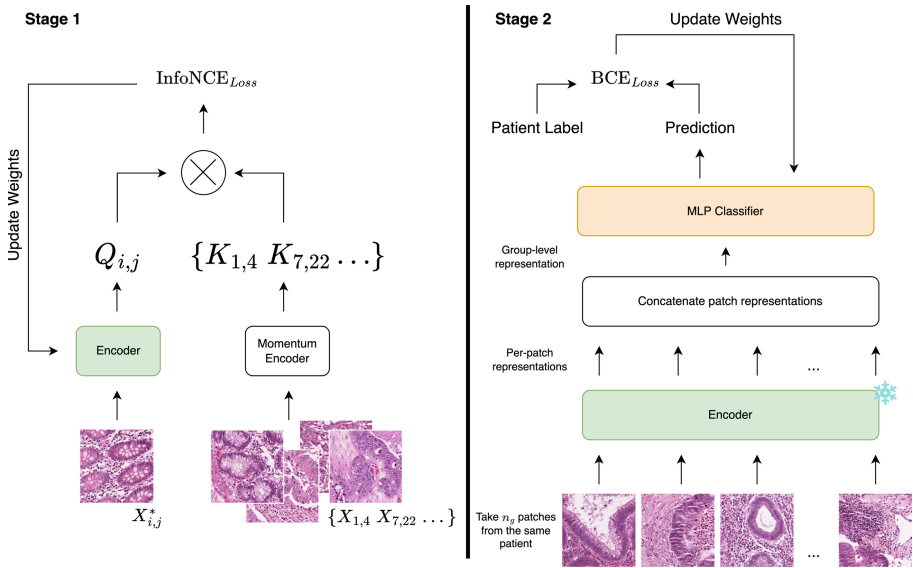


Fig. 2. Both stages of our proposed model.

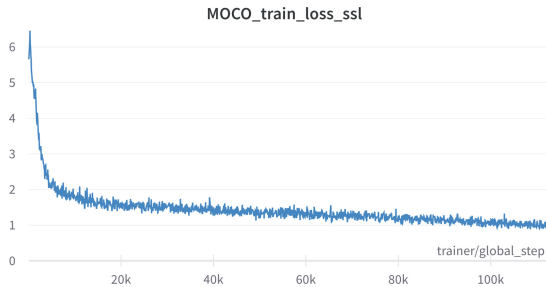
Stage 1 utilizes MoCo and stands to generate robust patch level embeddings that encode patches in a way such that they can be learned from in a downstream fashion. Above,  $Q_{i,j}$  represents the encoded query for patient  $i$  and patch  $j$ . Similarly  $K_{i,j}$  represents the same for patches encoded by the momentum encoder stored in the queue, and  $X_{i,j}$  corresponds to individual samples from the dataset relating to patient  $i$  and patch  $j$ . The stage 1 diagram is very similar to that found in Chen et al. [8].

Stage 2 groups the patches so that their features can be aggregated and the head of our model can learn from a set of patches as opposed to an individual sample. The encoder in stage 2 is a frozen copy of the trained encoder from stage 1. The snowflake indicates that its gradients are not tracked.

## 4.2 Stage 1: Training a Self-supervised Feature Extractor for Patch-Level Embeddings

In stage 1 of training, our feature extractor is trained in exactly the same way as described in the MoCo v2 paper [8]. Data loading and augmentation are unchanged. The main difference is our use of a Resnet18 [13] backbone as opposed to the Resnet50 (C4) backbone which was tested in the original implementation. This was done due to computational constraints and for comparison to the baseline approach from Kather. We also used cosine learning rate annealing, which seems to improve training. The output dimension of our feature extractor is 512 ( $n_o$ ).

To evaluate the ability of MoCo to extract usable features from patches, we tracked the value of the InfoNCE loss [20] on the training set. After achieving the lowest value for train InfoNCE loss (0.88 in our case), the model was saved and used for stage 2 training. This was not tracked on the validation dataset to avoid overfitting. Below you can see the training curve for MoCo over 621 epochs (Fig. 3):



**Fig. 3.** InfoNCE training loss over 621 epochs for MoCo.

The goal of contrastive learning, and thus momentum contrast learning, is to (as stated in the paper by He et al.) learn an encoder for a dictionary look up task. The contrastive loss function InfoNCE exists to return a low loss when encoded queries to the dictionary are similar to keys, and a high loss otherwise.

### 4.3 Stage 2: Training a Supervised Classifier on Patch Embedding Groups

In stage 2, the resnet18 [13] feature extractor trained by MOCO is frozen, and so gradients are not tracked. When making the forward pass, features extracted from patches are grouped by  $n_g$  (group size), meaning they are concatenated into one long vector. The length of this vector ( $l_g$ ) will be:

$$l_g = n_g * n_o \quad (3)$$

Meaning that the input dimension of our multi-layer perceptron (MLP) group-level classifier, or model head that we are training in stage 2 must have an input dimension of  $l_g$ . This brings us to the first issue regarding the  $n_g$  parameter. The larger the group size after feature extraction, the larger the first layer of the head must be. This is likely why we found a group size of 4 to be optimal for our dataset. When using a larger group size, the number of parameters for the head of the model increases dramatically, and it tends to overfit much faster.

As an interesting test, we also attempted  $n_g = 1$ , which performed very similarly to the standard approach. This is what we expected as it indicates the embedding space from Momentum Contrastive Learning is similarly effective to the embedding space of a model trained in a supervised fashion.

### 4.4 Evaluation

Judgement of the algorithm is performed using two main criteria. The first is patch level accuracy and the second is patient level accuracy.

Patch level accuracy ( $A_{patch}$ ) is exactly the same as accuracy in the general context. The only caveat is how the patches are assigned their label. Since our WSI's are labeled on a patient basis, the patches are labeled by inheriting the label of the patient to which they belong.

$$A_{patch} = \frac{\text{Number of Correctly Predicted Patches}}{\text{Total Number of Patches}} \quad (4)$$

Patient level accuracy ( $A_{patient}$ ) is a more crucial and more difficult to improve upon metric. It cannot be trained for directly, as an entire WSI cannot fit on GPU without downsampling. To measure this metric, we must save the models predictions on individual patches (or on groups of patches and extrapolate individual patch predictions) and calculate a final prediction for a patient



using the cumulative predictions of its constituent patches. This can be done using a majority vote approach or it is also possible to treat each patches prediction as a probability and average the probabilities before thresholding on a patient level and achieving a final prediction.

$$A_{patient} = \frac{\text{Number of Correctly Predicted Patients}}{\text{Total Number of Patients}} \tag{5}$$

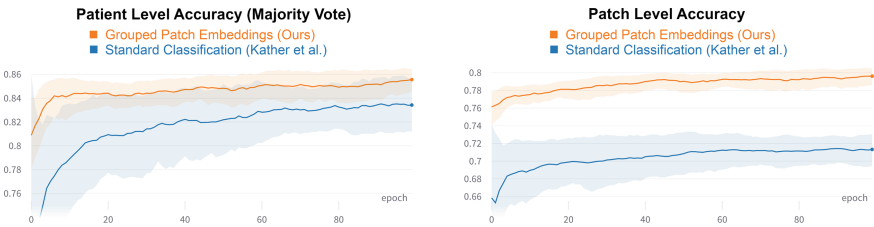
## 5 Experiments and Results

### 5.1 Standard Dataset

The results in this section refer to performance measured on the original dataset processed by Kather et al. [17]. We show the results of our implementation of Kather’s method compared to our improvement using momentum contrastive learning and group patch embeddings (Table 1 and Fig. 4).

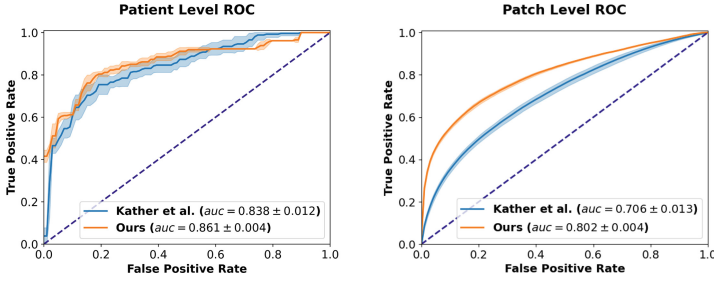
**Table 1.** Accuracy comparison on standard dataset. Both methods trained to 100 epochs. These are the validation results of an average of 10 runs per method. Our method achieves significantly higher accuracy in both patient (paired t-test,  $p < 0.001$ ), and patch level (paired t-test,  $p \ll 1e-6$ ) evaluation while also having a much more stable result given its smaller standard deviation.

Method	Patient accuracy	Patch accuracy
Ours	<b>0.862 ± 0.006</b>	<b>0.797 ± 0.005</b>
Kather et al.	0.837 ± 0.016	0.716 ± 0.015



**Fig. 4.** Validation accuracy during training. Average of 10 runs.

Due to the feature extractor already having been learned, our method initially trains much faster than the baseline. We have even noted that for some hyperparameter combinations it may be most effective to stop training after only a few epochs. And below are the ROC curves for the above models (Fig. 5):



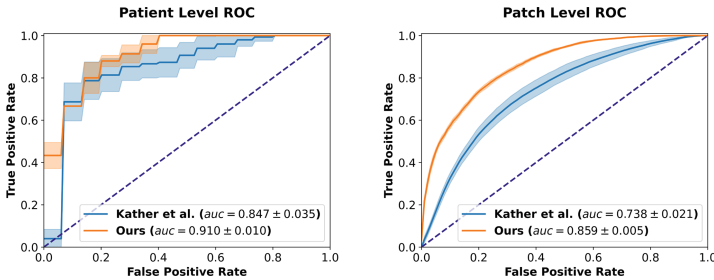
**Fig. 5.** ROC curves on standard validation dataset. Average of 10 runs. Our method achieves a significantly higher AUC in both patient (paired t-test,  $p < 0.01$ ) and patch level evaluation (paired t-test,  $p \ll 1e-7$ ).

## 5.2 Balanced Validation Set

The results in this section refer to performance measured on the balanced subset of the validation dataset processed by Kather et al. [17]. We describe the composition of this dataset in the second paragraph of the data section of this paper. Our method does even better on this balanced validation set compared to the original one from above. The differences were significant for both patient level (paired t-test,  $p \ll 1e-4$ ), and patch level (paired t-test,  $p \ll 1e-6$ ) classification. This suggest that our method is less prone to bias and overfitting to patients with more or less patches that have been extracted from them (Table 2 and Fig. 6).

**Table 2.** Accuracy comparison on balanced dataset.

Method	Patient accuracy	Patch accuracy
Ours	<b>0.797</b> $\pm$ 0.010	<b>0.751</b> $\pm$ 0.006
Kather et al.	0.723 $\pm$ 0.026	0.662 $\pm$ 0.013



**Fig. 6.** ROC curves on balanced dataset. Average of 10 runs. Our method achieves a significantly higher AUC in both patient (paired t-test,  $p < 0.01$ ) and patch level evaluation (paired t-test,  $p \ll 1e-6$ ).

## 6 Conclusions

Our work validates the feasibility of learning usable features from H&E stained biopsy slide patches using momentum contrast learning. We also qualify that learning from the aggregated features of multiple patches works better than simply averaging the predictions of individual patches for a WSI prediction. Finally, we contribute a simple and intuitive framework for combining these concepts with huge potential for improvement. The future for this domain lies in improving patch level feature extraction and aggregating more features to make global WSI decisions. The advent of a WSI classifier that is as accurate as laboratory testing for microsatellite status can drastically improve the rate at which patients are diagnosed and their treatment prospects.

## References

1. Bilal, M., et al.: Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit. Health* **3**, e763–e772 (2021)
2. Boland, C.R., Goel, A.: Microsatellite instability in colorectal cancer. *Gastroenterology* **138**(6), 2073–2087 (2010)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924 (2020)
4. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660 (2021)
5. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155 (2022)
6. Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. arXiv preprint [arXiv:2203.00585](https://arxiv.org/abs/2203.00585) (2022)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, pp. 1597–1607. PMLR (2020)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint [arXiv:2003.04297](https://arxiv.org/abs/2003.04297) (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE (2009)
10. Ehle, A., et al.: Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* **159**(4), 1406–1416 (2020)
11. Feng, Z., Xu, C., Tao, D.: Self-supervised representation learning by rotation feature decoupling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10364–10374 (2019)
12. Grill, J.B., et al.: Bootstrap your own latent—a new approach to self-supervised learning. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284 (2020)

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Hemati, S., Kalra, S., Meaney, C., Babaie, M., Ghodsi, A., Tizhoosh, H.: CNN and deep sets for end-to-end whole slide image representation learning. In: Medical Imaging with Deep Learning (2021)
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
16. Kather, J.N., et al.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**(7), 1054–1056 (2019)
17. Kather, J.: Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples. ZENODO (2019)
18. Li, K., Luo, H., Huang, L., Luo, H., Zhu, X.: Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int.* **20**(1), 1–13 (2020)
19. Liu, X., et al.: Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **35**, 857–876 (2021)
20. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748) (2018)
21. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: DeepSmile: self-supervised heterogeneity-aware multiple instance learning for dna damage response defect classification directly from H&E whole-slide images. arXiv preprint [arXiv:2107.09405](https://arxiv.org/abs/2107.09405) (2021)
22. Weinstein, J.N., et al.: The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**(10), 1113–1120 (2013)