SimTriplet: Simple Triplet Representation Learning with a Single GPU

Quan Liu[1], Peter C. Louis[2], Yuzhe Lu[1], Aadarsh Jha[1], Mengyang Zhao[1], Ruining Deng[1], Tianyuan Yao[1], Joseph T. Roland[2], Haichun Yang[2], Shilin Zhao[2], Lee E. Wheless[2], and Yuankai Huo[1]

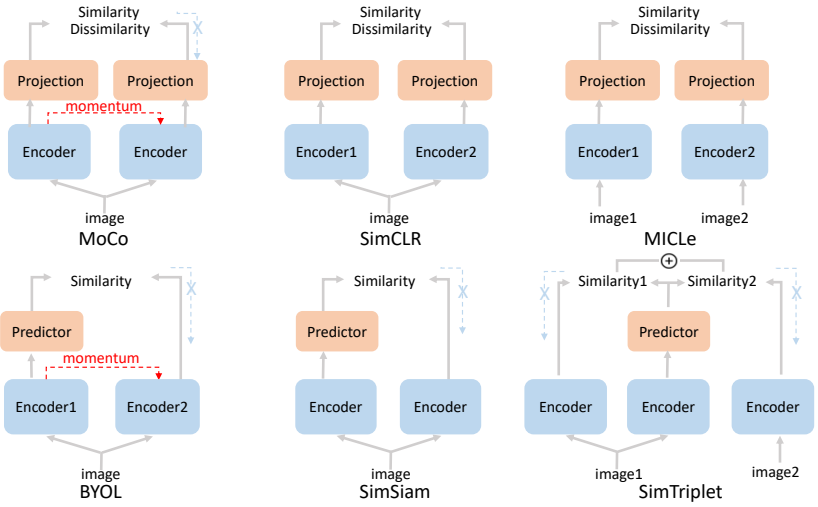
[1] Vanderbilt University, Nashville TN 37215, USA
[2] Vanderbilt University Medical Center, Nashville TN 37215, USA

Abstract. Contrastive learning is a key technique of modern self-supervised learning. The broader accessibility of earlier approaches is hindered by the need of heavy computational resources (e.g., at least 8 GPUs or 32 TPU cores), which accommodate for large-scale negative samples or momentum. The more recent SimSiam approach addresses such key limitations via stop-gradient without momentum encoders. In medical image analysis, multiple instances can be achieved from the same patient or tissue. Inspired by these advances, we propose a simple triplet representation learning (SimTriplet) approach on pathological images. The contribution of the paper is three-fold: (1) The proposed SimTriplet method takes advantage of the multi-view nature of medical images beyond self-augmentation; (2) The method maximizes both intra-sample and inter-sample similarities via triplets from positive pairs, without using negative samples; and (3) The recent mix precision training is employed to advance the training by only using a single GPU with 16GB memory. By learning from 79,000 unlabeled pathological patch images, SimTriplet achieved 10.58% better performance compared with supervised learning. It also achieved 2.13% better performance compared with SimSiam. Our proposed SimTriplet can achieve decent performance using only 1% labeled data. The code and data are available at https://github.com/hrlblab/SimTriplet.

Keywords: Contrastive learning · SimTriplet · Classification · Pathology.

1 Introduction

To extract clinically relevant information from GigaPixel histopathology images is essential in computer-assisted digital pathology [14, 21, 24]. For instance, the Convolutional Neural Network (CNN) based method has been applied to depreciate sub-tissue types on whole slide images (WSI) so as to alleviate tedious manual efforts for pathologists [22]. However, pixel-wise annotations are resource extensive given the high resolution of the pathological images. Thus, the fully supervised learning schemes might not be scalable for large-scale studies. To minimize the need of annotation, a well-accepted learning strategy is to first learn local image features through unsupervised



	MoCo	SimCLR	MICLe	BYOL	SimSiam	SimTriplet (ours)
No negative samples	✗	✗	✗	✓	✓	✓
Multiple views	✗	✗	✓	✗	✗	✓
No momentum	✗	✓	✓	✗	✓	✓
Computing resource (from original papers)	8~64 32GB GPUs	32~128 TPU cores	16~64 TPU cores	64 TPU v3 cores	1~8 GPUs	1 GPU

Fig. 1. Comparison of contrastive learning strategies. The upper panel compares the proposed SimTriplet with current representative contrastive learning strategies. The lower panel compares different approaches via a table.

feature learning, and then aggregate the features with multi-instance learning or supervised learning [13].

Recently, a new family of unsupervised representation learning, called contrastive learning (Fig. 1), shows its superior performance in various vision tasks [11,17,20,25]. Learning from large-scale unlabeled data, contrastive learning can learn discriminative features for downstream tasks. SimCLR [5] maximizes the similarity between images in the same category and repels representation of different category images. Wu et al. [20] uses an offline dictionary to store all data representation and randomly select training data to maximize negative pairs. MoCo [9] introduces a momentum design to maintain a negative sample pool instead of an offline dictionary. Such works demand large batch size to include sufficient negative samples (Fig. 1). To eliminate the needs of negative samples, BYOL [8] was proposed to train a model with a asynchronous momentum encoder. Recently, SimSiam [6] was proposed to further eliminate the momentum encoder in BYOL, allowing less GPU memory consumption.

To define different image patches as negative samples on pathological images is tricky since such a definition can depends on the patch size, rather than semantic differences. Therefore, it would be more proper to use nearby image patches as multi-view

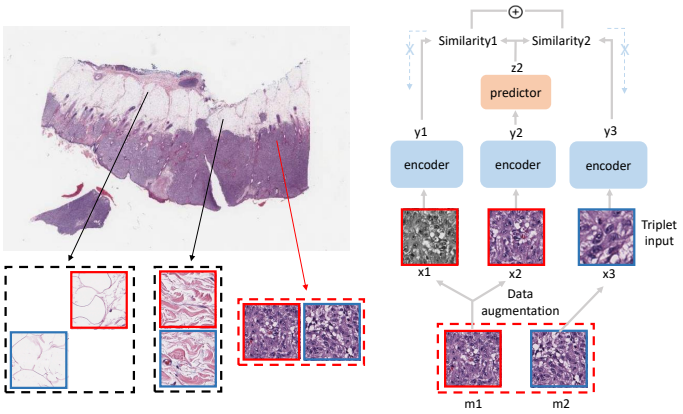


Fig. 2. Network structure of the proposed SimTriplet. Adjacent image pairs are sampled from unlabeled pathological images (left panel) for triplet representation learning (right panel). The GigaPixel pathological images provide large-scale "positive pairs" from nearby image patches for SimTriplet. Each triplet consists of two augmentation views from m_1 and one augmentation view from m_2. The final loss maximizes both inter-sample and intra-sample similarity as a representation learning.

samples (or called positive samples) of the same tissue type [19] rather than negative pairs. MICLe [2] applied multi-view contrastive learning to medical image analysis. Note that in [2, 19], the negative pairs are still needed within the SimCLR framework.

In this paper, we propose a simple triplet based representation learning approach (SimTriplet), taking advantage of the multi-view nature of pathological images, with effective learning by using only a single GPU with 16GB memory. We present a triplet similarity loss to maximize the similarity between two augmentation views of same image and between adjacent image patches. The contribution of this paper is three-fold:

• The proposed SimTriplet method takes advantage of the multi-view nature of medical images beyond self-augmentation.

• This method minimizes both intra-sample and inter-sample similarities from positive image pairs, without the needs of negative samples.

• The proposed method can be trained using a single GPU setting with 16GB memory, with batch size = 128 for 224×224 images, via mixed precision training.

2 Methods

The principle network of SimTriplet is presented in Fig 2. The original SimSiam network can be interpreted as an iterative process of two steps: (1) unsupervised clustering and (2) feature updates based on clustering (similar to K-means or EM algorithms) [6]. By knowing the pairwise information of nearby samples, the SimTriplet aims to further minimize the distance between the "positive pairs" (images from the same classes) in

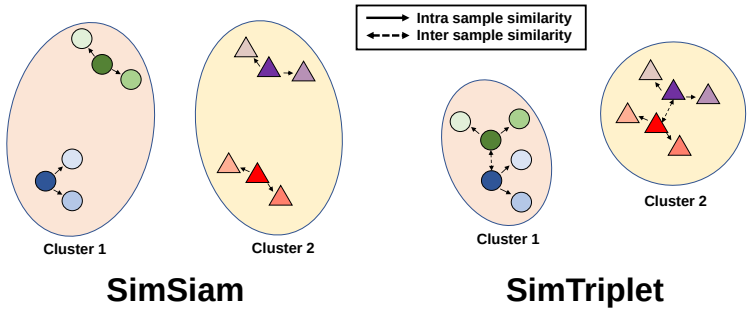


Fig. 3. Compare SimTriplet with SimSiam. SimSiam network maximizes intra-sample similarity by minimizing distance between two augmentation views from the same image. The proposed SimTriplet model further enforce the inter-sample similarity from positive sample pairs.

the embedding space (Fig. 3). In the single GPU setting with batch size 128, SimTriplet provides more rich information for the clustering stage.

2.1 Multi-view nature of medical images

In many medical image analysis tasks, multi-view (or called multi-instance) imaging samples from the same patient or the same tissue can provide complementary representation information. For pathological images, the nearby image patches are more likely belong to the same tissue type. Thus, the spatial neighbourhood on a WSI provide rich "positive pairs" (patches with same tissue types) for triplet representation learning. Different from [12], all samples in our triplets are positive samples, inspired by [6]. To train SimTriplet, we randomly sample image patches as well as their adjacent patches (from one of eight nearby locations randomly) as positive sample pairs from the same tissue type.

2.2 Triplet representation learning

Our SimTriplet network forms a triplet from three randomly augmented views by sampling positive image pairs (Fig. 2). The three augmented views are fed into the encoder network. The encoder network consists of a backbone network (ResNet-50 [10]) and a three-layer multi-layer perceptron (MLP) projection header. The three forward encoding streams share the same parameters. Next, an MLP predictor is used in the middle path. The predictor processes the encoder output from one image view to match with the encoder output of two other image views. We applies stop-gradient operations to two side paths. When computing loss between predictor output and image representation from encoder output, encoded representation is regarded as constant [6]. Two encoders on side paths will not be updated by back propagation. We used negative cosine similarity Eq.(1) between different augmentation views of (1) the same image patches, and (2) adjacent image patches as our loss function. For example, image m_1 and image m_2

are two adjacent patches cropped from the original whole slide image (WSI). x_1 and x_2 are randomly augmented views of image m_1, while x_3 is the augmented view of image m_2. Representation y_1, y_2 and y_3 are encoded from augmented views by encoder. z_1, z_2 and z_3 are the representation processed by the predictor.

$$\mathcal{C}(p,q) = -\frac{p}{\|p\|_2} \cdot \frac{q}{\|q\|_2} \tag{1}$$

$\mathcal{L}_{Intrasample}$ is the loss function to measure the similarities between two augmentation views x_1 and x_2 of image m_1 as seen in Eq.(2).

$$\mathcal{L}_{Intrasample} = \frac{1}{2}\mathcal{C}(y_1, z_2) + \frac{1}{2}\mathcal{C}(y_2, z_1) \tag{2}$$

$\mathcal{L}_{Intersample}$ is the loss function to measure the similarities between two augmentation views x_2 and x_3 of adjacent image pair m_1 and m_2 as in Eq.(3).

$$\mathcal{L}_{Intersample} = \frac{1}{2}\mathcal{C}(y_2, z_3) + \frac{1}{2}\mathcal{C}(y_3, z_2) \tag{3}$$

The triplet loss function as used in our SimTriplet network is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{Intrasample} + \mathcal{L}_{Intersample} \tag{4}$$

$\mathcal{L}_{Intrasample}$ minimizes the distance between different augmentations from the same image. $\mathcal{L}_{Intersample}$ minimizes the difference between nearby image patches.

2.3 Expand batch size via mix precision training

Mix precision training [16] was invented to offer significant computational speedup and less GPU memory consumption by performing operations in half-precision format. The minimal information is stored in single-precision to retain the critical parts of the training. By implementing the mix precision to SimTriplet, we can extend the batch size from 64 to 128 to train images with 224×224 pixels, using a single GPU with 16GB memory. The batch size 128 is regarded as a decent batch size in SimSiam [6].

3 Data and Experiments

3.1 Data

Annotated data. We extracted image patches from seven melanoma skin cancer Whole Slide Images (WSIs) from the Cancer Genome Atlas (TCGA) Datasets (TCGA Research Network: https://www.cancer.gov/tcga). From the seven annotated WSIs, 4698 images from 5 WSIs were obtained for training and validation, while 1,921 images from 2 WSIs were used for testing. Eight tissue types were annotated as : blood vessel (353 train 154 test), epidermis (764 train 429 test), fat (403 train 137 test), immune cell (168 train 112 test), nerve (171 train 0 test), stroma (865 train 265 test), tumor (1,083 train 440 test) and ulceration (341 train 184 test).

Following [18, 23]), each image was a 512×512 patch extracted from 40× magnification of a WSI with original pixel resolution 0.25 micron meter. The cropped image samples were annotated by a board-certified dermatologist and confirmed by another pathologist. Then, the image patches were resized to 128×128 pixels. Note that the 224×224 image resolution provided 1.8% higher balance accuracy (based on our experiments) using the supervised learning. We chose 128×128 resolution for all experiments for a faster training speed.

Unlabeled data. Beyond the 7 annotated WSIs, additional 79 WSIs without annotations were used for training contrastive learning models. The 79 WSIs were all available and usable melanoma cases from TCGA. The number and size of image patches used for different contrastive learning strategies are described in §**Experiment**.

3.2 Supervised learning

We used ResNet-50 as the backbone in supervised training, where the optimizer is Stochastic Gradient Descent (SGD) [3] with the base learning rate $lr = 0.05$. The optimizer learning rate followed (linear scaling [7]) $lr \times \text{BatchSize}/256$. We used 5-fold cross validation by using images from four WSIs for training and image from the remaining WSI for validation. We trained 100 epochs and selected best model based on validation. When applying the trained model on testing images, the predicted probabilities from five models were averaged. Then, the class with the largest ensemble probability was used as the predicted label.

3.3 Training contrastive learning benchmarks

We used the SimSiam network [5] as the baseline method of contrastive learning. Two random augmentations from the same image were used as training data. In all of our self-supervised pre-training, images for model training were resized to 128×128 pixels. We used momentum SGD as the optimizer. Weight decay was set to 0.0001. Base learning rate was $lr = 0.05$ and batch size equals 128. Learning rate was $lr \times \text{BatchSize}/256$, which followed a cosine decay schedule [15]. Experiments were achieved only on a single GPU with 16GB memory. Models were pre-trained for $39,500/128 \times 400 \approx 127,438$ iterations. 79 unlabeled WSIs were used for self-supervised pre-training. We randomly cropped 500 images from each WSI and resized them to 128×128 pixels. 39,500 images in total serve as the original training data.

Following MICLe [2], we employed multi-view images as two inputs of the network. Since we did not use negative samples, multi-view images was trained by SimSiam network instead of SimCLR. For each image in the original training dataset, we cropped one patch which is randomly selected from its eight adjacent patches consisting of an adjacent images pairs. We had 79,000 images (39,500 adjacent pairs) as training data. Different from original SimSiam, network inputs were augmentation views of an adjacent pair. Referring to [6], we applied our data on SimSiam network. First, we used 39,500 images in original training dataset to pre-train on SimSiam. To see the impact of training dataset size, we randomly cropped another 39,500 images from 79 WSIs for training on a larger dataset of 79,000 images. We then used training data from the MICLe experiment to train the SimSiam network.

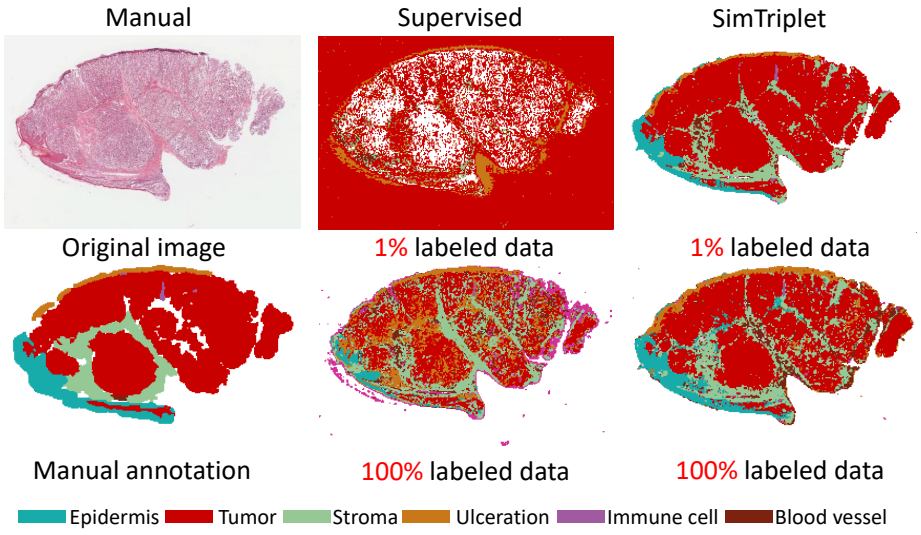


Fig. 4. Visualization of classification results. One tissue sample is manually segmented by our dermatologist (via QuPath software) to visually compare the classification results. The contrasting learning achieved superior performance compared with supervised learning, even using only 1% of all available labeled data.

3.4 Training the proposed SimTriplet

The same 79,000 images (39,500 adjacent pairs) were used to train the SimTriplet. Three augmentation views from each adjacent pair were used as network inputs. Two augmentation views were from one image, while the other augmentation view was augmented from adjacent images. Three augmentation views were generated randomly, where the augmentation settings were similar with the experiment on SimSiam [5]. Batch size was 128 and experiment run on a single 16GB memory GPU.

3.5 Linear evaluation (fine tuning)

To apply the self-supervised pre-training networks, as a common practice, we froze the pretrained ResNet-50 model by adding one extra linear layer which followed the global average pooling layer. When finetuning with the annotated data, only the extra linear layer was trained. We used the SGD optimizer to train linear classifier with a based (initial) learning rate lr=30, weight decay=0, momentum=0.9, and batch size=64 (follows [6]). The same annotated dataset were used to finetune the contrastive learning models as well as to train supervised learning. Briefly, 4,968 images from 5 annotated WSIs were divided into 5 folders. We used 5-fold cross validation: using four of five folders as training data and the other folder as validation. We trained linear classifiers for 30 epochs and selected the best model based on the validation set. The pretrained

models were applied to the testing dataset (1,921 images from two WSIs). As a multi-class setting, macro-level average F1 score was used [1]. Balanced accuracy was also broadly used to show the model performance on unbalanced data [4].

4 Results

Model classification performance. F1 score and balanced accuracy were used to evaluate different methods as described above. We trained a supervised learning models as the baseline. From Table 1, our proposed SimTriplet network achieved the best performance compared with the supervised model and SimSiam network [6] with same number of iterations. To show a qualitative result, a segmentation of a WSI from test dataset is shown in Fig. 4.

Table 1. Classification performance.

Methods	Unlabeled Images	Paired Inputs	F1 Score	Balanced Acc
Supervised	0		0.5146	0.6113
MICLe [2]*	79k	✓	0.5856	0.6666
SimSiam [6]	39.5k		0.5421	0.5735
SimSiam [6]	79k	✓	0.6267	0.6988
SimSiam [6]	79k		0.6275	0.6958
SimTriplet (ours)	79k	✓	**0.6477**	**0.7171**

* We replace SimCLR with SimSiam.

Model performance on partial training data. To evaluate the impact of training data number, we trained a supervised model and fine-tuned a classifier of the contrastive learning model on different percentages of annotated training data (Table 2). Note that for 1% to 25%, we ensure different classes contribute a similar numbers images to address the issue that the annotation is highly imbalanced.

Table 2. Balanced Acc of using different percentage of annotated data.

Methods	Percentage of Used Annotated Training Data			
	1%	10%	25%	100%
Supervised	0.0614	0.3561	0.4895	0.6113
SimSiam [6]	0.7085	0.6864	0.6986	0.6958
SimTriplet	**0.7090**	**0.7110**	**0.7280**	**0.7171**

5 Conclusion

In this paper, we proposed a simple contrastive representation learning approach, named SimTriplet, advanced by the multi-view nature of medical images. Our proposed contrastive learning methods maximize the similarity between both self augmentation views and pairwise image views from triplets. Moreover, our model can be efficiently trained on a single GPU with 16 GB memory. The performance of different learning schemes are evaluated on WSIs, with large-scale unlabeled samples. The proposed SimTriplet achieved superior performance compared with benchmarks, including supervised learning baseline and SimSiam method. The contrastive learning strategies showed strong generalizability by achieving decent performance by only using 1% labeled data.

References

1. Attia, M., Samih, Y., Elkahky, A., Kallmeyer, L.: Multilingual multi-class sentiment classification using convolutional neural networks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
2. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M.: Big self-supervised models advance medical image classification (2021)
3. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. pp. 3121–3124 (2010). https://doi.org/10.1109/ICPR.2010.764
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566 (2020)
7. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
8. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
10. 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)
11. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
13. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

14. Liskowski, P., Krawiec, K.: Segmenting retinal blood vessels with deep neural networks. IEEE transactions on medical imaging **35**(11), 2369–2380 (2016)
15. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017)
16. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training (2018)
17. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
18. Raju, A., Yao, J., Haq, M.M., Jonnagaddala, J., Huang, J.: Graph attention multi-instance learning for accurate colorectal cancer staging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–539. Springer (2020)
19. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
20. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination (2018)
21. Xu, Y., Jia, Z., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C.: Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 947–951. IEEE (2015)
22. Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C.: Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC bioinformatics **18**(1), 1–17 (2017)
23. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4837–4846 (2020)
24. Zhu, C., Zou, B., Zhao, R., Cui, J., Duan, X., Chen, Z., Liang, Y.: Retinal vessel segmentation in colour fundus images using extreme learning machine. Computerized Medical Imaging and Graphics **55**, 68–77 (2017)
25. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6002–6012 (2019)