Convolutional Masked Image Modeling for Dense Prediction Tasks on Pathology Images

Yan Yang[1] Liyuan Pan[2] Liu Liu[3] Eric A. Stone[1] †
[1]BDSI, ANU [2]BITSZ & School of CSAT, BIT [3]Cyberverse Lab
{yan.yang, eric.stone}@anu.edu.au liyuan.pan@bit.edu.cn nwpuliuliu@gmail.com

Abstract

This paper studies a convolutional masked image modeling approach for boosting downstream dense prediction tasks on pathology images. Our method is self-supervised, and entails two strategies in sequence. Considering features contained in the pathology images usually have a large spatial span, e.g.*, glands, we insert* `[MASK]` *tokens to the masked regions after the stem layer of the convolutional network for encoding unmasked pixels, which facilitates information propagation through masked regions for reconstructing unmasked pixels. Furthermore, the pathology images contain features that are represented in diverse affine shapes and color spaces. We, therefore, enforce the network to learn the affine and color invariant embedding by imposing transformation constraints between the unmasked image-encoded embedding and reconstruction targets. Our approach is simple but effective. With extensive experiments on standard benchmark datasets, we demonstrate superior transfer learning performance on downstream tasks over past state-of-the-art approaches.*

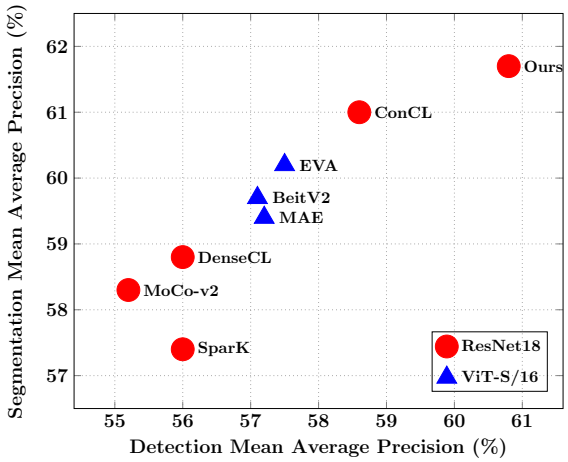


Figure 1. *Overall comparison of our method and state-of-the-art self-supervised learning framework on GlaS dataset [41] for detection and instance segmentation tasks. The triangles and cycles respectively denote the ViT-S/16 and ResNet18 [22] backbones of self-supervised learning methods. The mean average precision (%) [34, 47] is used as the evaluation metric. We show the top-performing methods only, and please refer to our experiment section (Tab. 1) for details. Best viewed in color on the screen.*

1. Introduction

Deep learning on computational pathology has shown promising trends in precise and efficient diagnosis, prognosis, and treatment selection by using pathology images, i. e., whole slide images [29, 35, 42, 49]. However, the performance of the deep learning methods is hindered by limited amounts of dense manual annotations of pathology images that are composed of gigapixels.

To mitigate the annotation workload, this paper studies a self-supervised pre-training framework on pathology images. Our method is a convolutional masked image modeling approach, only using pathology images for training. By transferring our pre-trained network weights into diverse computational pathology downstream tasks, significant performance improvement can be achieved, compared with training the network from scratch.

Self-supervised learning (SSL) has extraordinarily succeeded on natural RGB images, shifting state-of-the-art approaches from adopting contrastive learning to emerging masked image modeling (MIM). However, the MIM approaches [43] perform inferior over contrastive learning approaches [47] in pathology images (Fig. 1).

A question is naturally raised to our study, *what is obstructing the success of MIM for SSL on pathology images?* In the following, with detailed analysis, we consider convolution network backbones [22] for pre-training to assist the victory of MIM, due to their high performance on downstream tasks over vision transformer (ViT) backbones [14] (refer to Tab. 1 for evidence).

Under the framework of MIM, networks are trained to reconstruct masked pixels from unmasked ones. Past con-

† Corresponding author.

volutional MIM approaches use sparse convolutions [9, 43] for encoding unmasked pixels into embeddings, `[MASK]` tokens are then inserted into masked regions of the embeddings, and they are finally used to decode the masked pixels.

Therefore, compared to contrastive learning, convolutional MIM suffers two limitations: **i)** *information propagation bottlenecks*. Features contained in the pathology images usually have a large spatial span (Fig. 2). Though unmasked pixels are removed by applying sparse convolutions, the information propagation among the unmasked pixels is also hindered during encoding, which is in contrast to contrastive learning that allows information propagation in any image region typically. In the worst case, a pathology image is masked into isolated blocks, and the information propagation among the blocks only happens in BatchNorm layers and other bottom layers (when the sparse convolution kernel size is larger than the distance among the isolated blocks after progressive downsampling operations) of the network; **ii)** *lack of embedding invariance*. Pathology images contain features represented in diverse affine shapes and color space [23, 25]. Unlike contrastive learning that the networks are trained to learn invariant embedding from differently represented same pathology image features, networks trained under the MIM framework encode feature-variant embedding, for the pathology image reconstruction. However, the invariant embedding is demanded by diverse pathology downstream dense prediction tasks.

Our convolutional MIM method proposes two strategies to mitigate the above limitations in sequence. Given a pathology image and a randomly generated mask, we first perform affine and color augmentations on the image. Instead of using sparse convolutions for encoding the unmasked pixels, our masking operation has two stages. We zero out masked pixels, forward the masked images to the stem layer of the network, and then insert `[MASK]` tokens to the stem layer output before further network encoding, allowing information propagation of unmasked pixels through masked regions for the network.

After encoding unmasked pixels to embeddings, the inverse affine transformations are performed to the embeddings, for reconstructing corresponding unmasked pixels in the un-augmented pathology images. This is inspired by contrastive learning that networks are trained to learn affine and color invariant embeddings. In this way, our network is trained to learn the same embedding for pathology image features that are even affine transformed and represented in different colors, benefiting potential downstream tasks.

Our contributions are summarised as follows:

- A convolutional masked image modeling framework for self-supervised learning on pathology images;
- A masking strategy, avoiding information propagation bottleneck during encoding unmasked pixels;
- An invariant embedding learning strategy by constraining the network reconstruction targets.

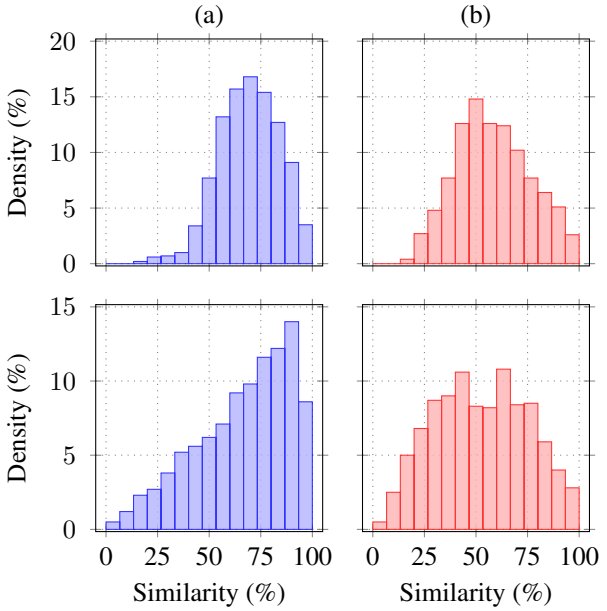


Figure 2. *Region similarity distribution of (a) pathology images and (b) natural images. In the first and second rows, we respectively use the large-scale IN-1K [12] and multi-source pathology datasets [1] pre-trained ResNet18 [1, 22] for extracting spatial features (i. e., output of the conv5 layer) of 500 randomly sampled (a) pathology images [26] and (b) natural images [12], and calculate cosine similarity between different regions. Considering region similarity within natural images as a reference, the comparison shows that regions of pathology image share high similarity that tends to compose features of a large spatial span.*

We evaluate our method on standard pathology image benchmark datasets. Comparing with extensive SSL approaches and recently emerged vision foundation models (i. e., SAM [27]), we show that our method achieves state-of-the-art performance. Refer to Fig. 1 for an overall comparison with state-of-the-art approaches.

2. Related Work

SSL frameworks are mainly underpinned by two main approaches: contrastive learning and masked image modeling. In this section, we introduce recent SSL achievements, and then review SSL applications in pathology images.

Contrastive learning. Learning invariant image embeddings is focused on contrastive learning approaches [5]. By assuming augmentation invariance of images, multiple views of each image are generated for instance discrimination [5, 6, 8, 20, 46] or group discrimination [3, 4, 30, 47]. This pulls the embeddings of positively matched embedding pairs *e.g.*, views of the same images, while pushing away the negatively matched embedding pair *e.g.*, views of different images. There are also some approaches that

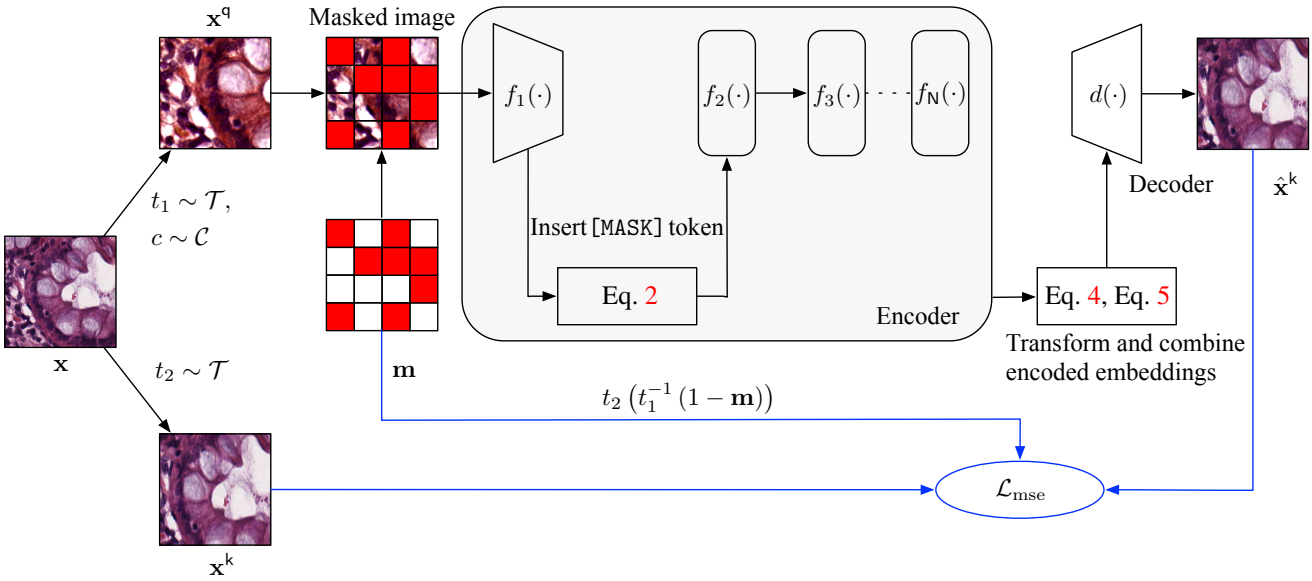


Figure 3. *Overall architecture. Given a pathology slide image* $\mathbf{x}$*, we sample two affine transformation,* $t_1(\cdot), t_2(\cdot) \sim \mathcal{T}$*, and a color transformation,* $c \sim \mathcal{C}$*, where* $\mathcal{T}$ *and* $\mathcal{C}$ *are respectively affine and color transformation space. We then transform the image* $\mathbf{x}$ *to* $\mathbf{x}^{\mathsf{q}} = c\left(t_1\left(\mathbf{x}\right)\right)$ *and* $\mathbf{x}^{\mathsf{k}} = t_2(\mathbf{x})$*. Meanwhile, we randomly sample a mask* m*, where the masked regions are colored in red. We denote our encoder network as a* N*-layer sequence,* $f_1 \circ f_2 \circ \cdots \circ f_{\mathsf{N}}(\cdot)$*. We first mask the image* $\mathbf{x}^{\mathsf{q}}$ *by a elementwise multiplication* $\odot$ *with* $\mathbf{m}$*, embed the masked image* $\mathbf{x}^{\mathsf{q}} \odot \mathbf{m}$ *with the first/stem layer* $f_1(\cdot)$*, and insert the* `[MASK]` *token (Eq. 2) to the masked regions of the stem layer output* $f_1(\mathbf{x}^{\mathsf{q}} \odot \mathbf{m})$ *based on the mask* $\mathbf{m}$ *before forwarding the embeddings into further encoder layers. After the encoding stage, we collect intermediate embeddings from each encoder layer, transform each of the embeddings with* $t_2\left(t_1^{-1}(\cdot)\right)$*, and combine the transformed embeddings (Eq. 4 and Eq. 5) to construct a simple multi-scale feature, for reconstructing* $\mathbf{x}^{\mathsf{k}}$ *with a decoder* $d(\cdot)$*. We denote the reconstruction as* $\hat{\mathbf{x}}^{\mathsf{k}}$*. We optimize the network by the mean squared loss* $\mathcal{L}_{mse}$ *(Eq. 7), encouraging the network to reconstruct the corresponding masked regions* $t_2\left(t_1^{-1}(1-\mathbf{m})\right)$ *of* $\mathbf{x}^{\mathsf{k}}$*. Best viewed in color on the screen.*

only enforce embedding similarity [7, 17]. In certain tasks, contrastive learning has shown remarkable performance improvements compared to supervised pre-training [8, 14, 48]. However, the instance and group discrimination settings are biased in learning image-level embeddings, and are less effective for dense prediction tasks. Therefore, the instance discrimination tasks are further refined into discriminating embeddings of different regions [16, 24, 39, 40, 44], forming dense contrastive learning. Motivated by contrastive learning, this paper studies the idea of learning invariant embedding in a MIM manner.

Masked image modeling. Inspired by masked language modeling [13], MIM trains a vision transformer [14] to reconstruct masked pixels from unmasked ones [19], where the masking operations are usually performed patch-wisely. By leveraging vision foundation models [37], the reconstruction targets can be represented by the foundation model embeddings, learning stronger structure and semantic information of the images. Some works also frame the reconstruction tasks into a classification problem, by predicting the indices assigned to the masked patches by an image tokenizer [2, 36, 50]. One may note that MIM brings irregularly masked images, imposing challenges for applications in regular convolutional networks. Therefore, past works [43] apply sparse convolutions that treat unmasked patches as sparse voxels, showing state-of-the-art transfer learning performance on diverse downstream tasks. This paper studies convolutional MIM without relying on sparse convolutions, by proposing a new masking strategy.

Self-supervised learning in pathology images. Explorations of SSL in pathology images are still rare, and can be specified to [10, 28, 47], where [10] applies SimCLR [5] in pathology images, [28] explores pathology-related auxiliary tasks for pre-training, *e.g.*, magnification prediction, and [47] proposes a region-based contrastive learning approach. In the same vein of [47], this paper studies SSL in pathology images for dense prediction tasks, but focuses on learning discriminative network embeddings through MIM.

3. Method

We start with introducing the background knowledge. Then, we provide an overview of our convolutional MIM framework, and describe each network component. Our framework is shown in Fig. 3.

Preliminary. Convolutional MIM follows a pipeline of learning discriminative embeddings through reconstructing masked images (refer to [43] for details). We briefly re-

view the main steps. An image $\mathbf{x}$ is augmented to $\mathbf{x}^{\mathsf{q}}$, and a mask $\mathbf{m}$ is randomly generated, where values of one and zero of the mask respectively denote unmasked and masked regions The masked image $\mathbf{x}_{\mathsf{q}} \odot \mathbf{m}$ is next encoded with a sparse convolution-based network, where $\odot$ is an element-wise multiplication operator. After the encoding stage, the sparse embeddings from each network layer are collected and densified. Specifically, for embedding from each network layer, a mask `[MASK]` token is broadcasted to the same spatial size of the embedding, and then inserted into masked regions of the embedding according to the mask $\mathbf{m}$. These densified embeddings are finally forwarded to a decoder for reconstructing the masked pixels $\mathbf{x}^{\mathsf{q}} \odot (1 - \mathbf{m})$.

Two main limitations exist in the past convolutional MIM for applying in pathology images, i. e., information propagation bottlenecks and lack of embedding invariance, as analyzed in Sec. 1. In this study, we attempt to mitigate the limitations by rethinking the masking strategy and reconstruction targets.

Model overview. Given a image $\mathbf{x}$, we sample two affine transformation, $t_1(\cdot) \sim \mathcal{T}$ and $t_2(\cdot) \sim \mathcal{T}$, and a color transformation, $c(\cdot) \sim \mathcal{C}$, and augment the image by $\mathbf{x}^{\mathsf{q}} = c(t_1(\mathbf{x}))$, where $\mathcal{T}$ and $\mathcal{C}$ are space of affine and color transformations. Our method has three components connected in sequence: i) embedding encoding. With a randomly generated mask $\mathbf{m}$, we use regular convolutions for encoding masked image $\mathbf{x}_{\mathsf{q}} \odot \mathbf{m}$, where a `[MASK]` token is inserted to the intermediate network embedding outputted from the stem layer of the network; ii) embedding decoding. Collecting embeddings from each layer of the network, we perform affine transformation $t_2\left(t_1^{-1}(\cdot)\right)$, aiming to decode the unmasked pixels of $\mathbf{x}^{\mathsf{k}} = t_2(\mathbf{x})$, i. e., $t_2\left((1-\mathbf{m}) \odot \mathbf{x}\right)$ or $t_2\left(1-\mathbf{m}\right) \odot \mathbf{x}^{\mathsf{k}}$; iii) loss. We optimize our network (Fig. 3) with mean squared error (MSE), following [19, 43].

Embedding encoding. We denote our N-layer backbone network as $f_1 \circ f_2 \circ \cdots \circ f_{\mathsf{N}}(\cdot)$[1], where $\mathbf{f}_i$ is the i^{th} layer and $1 \leq i \leq \mathsf{N}$. We forward the masked image $\mathbf{x}^{\mathsf{q}} \odot \mathbf{m}$ to the stem layer $f_1(\cdot)$, for embedding the masked image,

$$\mathbf{e} = f_1(\mathbf{x}^{\mathsf{q}} \odot \mathbf{m}) \,, \tag{1}$$

where $\mathbf{e}$ is the embedding from the stem layer.

Unlike the past convolutional MIM that encodes the images with sparse convolutions and inserts `[MASK]` tokens to the encoded embeddings for decoding masked pixels, we insert the `[MASK]` tokens to $\mathbf{e}$ during encoding, enjoying information propagation across different image regions, under a regular convolution network. Specifically, we first broadcast the `[MASK]` tokens to $\texttt{[MASK]}^{\mathsf{stem}}$, and the mask $\mathbf{m}$ is also interpolated to $\mathbf{m}^{\mathsf{stem}}$, for having the same spatial size of $\mathbf{e}$. We then insert the $\texttt{[MASK]}^{\mathsf{stem}}$ tokens to $\mathbf{e}$ with $\mathbf{m}^{\mathsf{stem}}$, yielding $\mathbf{z}_1$. The process is given by

$$\mathbf{z}_1 = \mathbf{e} \cdot \mathbf{m}^{\mathsf{stem}} + (1 - \mathbf{m}^{\mathsf{stem}}) \cdot \texttt{[MASK]}^{\mathsf{stem}} \,. \tag{2}$$

To further encode $\mathbf{z}_1$, the remaining network layers $\{f_i(\cdot)\}_{i=2}^{\mathsf{N}}$ progressively forward the embeddings,

$$\mathbf{z}_i = f_i(\mathbf{z}_{i-1}) \,, \quad \forall i \in [2, \cdots, \mathsf{N}] \,. \tag{3}$$

Collecting embedding $\mathbf{z}_i$ from each layer, we have $\{\mathbf{z}_i\}_{i=1}^{\mathsf{N}}$, preparing for decoding masked pixels.

Embedding decoding. To learn invariant features, we perform an inverse affine transformation $t_1^{-1}(\cdot)$ on the collected embeddings $\{\mathbf{z}_i\}_{i=1}^{\mathsf{N}}$, for aligning the embeddings with the un-augmented version of $\mathbf{x}^{\mathsf{q}}$, i. e., $\mathbf{x}$. We then further augment the inverse transformed embedding with $t_2(\cdot)$, for reconstructing unmasked pixels of $\mathbf{x}^{\mathsf{k}}$, i. e., $t_2(\mathbf{x})$. The overall transformation process is given by

$$\mathbf{z}_i^{\mathsf{k}} = t_2\left(t_1^{-1}\left(\mathbf{z}_i\right)\right) \,, \quad \forall i \in [1, \cdots, \mathsf{N}] \,. \tag{4}$$

Note that the network is encouraged to learn color invariant embedding by encouraging to reconstruct corresponding masked pixels of $\mathbf{x}^{\mathsf{k}}$ from unmasked pixels of $\mathbf{x}^{\mathsf{q}}$.

To reconstruct the masked pixels, we first unify the embeddings $\{\mathbf{z}_i^{\mathsf{k}}\}_{i=1}^{\mathsf{N}}$ to the same spatial size by using N convolution layers $\{\mathrm{Conv}_i(\cdot)\}_{i=1}^{\mathsf{N}}$, and combine the outputs to construct a simple multi-scale embedding $\mathbf{z}^{\mathsf{msc}}$,

$$\mathbf{z}^{\mathsf{msc}} = \sum_{i=1}^{\mathsf{N}} \mathrm{Conv}_i\left(\mathbf{z}_i^{\mathsf{k}}\right) \,. \tag{5}$$

With $\mathbf{z}^{\mathsf{msc}}$, a decoder $d(\cdot)$ is used for reconstructing $\mathbf{x}^{\mathsf{k}}$. We denote the reconstruction as $\hat{\mathbf{x}}^{\mathsf{k}}$ that is given by

$$\hat{\mathbf{x}}^{\mathsf{k}} = d(\mathbf{z}^{\mathsf{msc}}) \,. \tag{6}$$

Loss. Our network is optimized with MSE loss for pulling the reconstruction of masked pixels close to the target image $\mathbf{x}^{\mathsf{k}} = t_2(\mathbf{x})$. The loss is given by

$$\mathcal{L}_{\mathrm{mse}} = \left\| \left(\mathbf{x}^{\mathsf{k}} - \hat{\mathbf{x}}^{\mathsf{k}}\right) \odot t_2\left(t_1^{-1}\left(1 - \mathbf{m}\right)\right) \right\|^2 \,, \tag{7}$$

where $\|\cdot\|$ is the Frobenius norm that reduces the loss from a matrix to a scalar.

4. Experiment

4.1. Experimental Setup

Pre-training dataset. We use the NCT-CRC-HE-100K-NONORM (NCT) dataset [26] for pre-training. The dataset contains 1×10^6 pathology image tiles extracted from hematoxylin & eosin (H&E) stained whole slide images (WSIs) of human colorectal cancer and normal tissue. They are recorded at 0.5 microns per pixel, and each pathology image tile is composed of 224×224 pixels. We follow the train-test splits of [47], and use the train split for network pre-training.

[1]The first layer $f_1(\cdot)$ of a convolutional network is usually known as the stem layer. Similar to ViT [14], it can be considered for the purpose of embedding the input, i. e., representing the input in high dimensional feature space. For example, the first/stem layer of ResNet [22].

Table 1. *Comparison of object detection and instance segmentation of our method and state-of-the-art approaches on the GlaS dataset [41] and CRAG dataset [18]. We use AP^{bb}, AP^{bb}_{75}, AP^{mk}, and AP^{mk}_{75} as evaluation metrics [34]. AP^{bb} and AP^{mk} respectively measure the mean average bounding box and segmentation precision, while AP^{bb}_{75} and AP^{mk}_{75} are the precision at 75% intersection over union (IoU).*

Method	Backbone	GlaS dataset				CRAG dataset			
		Detection		Segmentation		Detection		Segmentation	
		AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}
Training from scratch, i. e., random weight initilizations.									
ResNet18	ResNet18	49.8	57.3	52.1	60.7	51.1	57.0	50.6	57.3
ViT-S/16	ViT-S/16	49.3	55.5	51.2	58.3	57.6	52.5	48.8	52.7
Supervised pre-training, i. e., the backbone weights are transferred from supervised pre-training on the NCT dataset.									
ResNet18	ResNet18	50.2	56.9	53.2	62.1	49.2	55.2	49.4	55.0
ViT-S/16	ViT-S/16	44.2	49.4	47.3	53.4	40.6	43.1	42.1	44.2
Contrastive learning, i. e., the backbone weights are transferred from contrastive learning.									
SimCLR	ResNet18	50.6	56.8	53.9	62.9	48.1	52.0	47.8	53.7
BYOL	ResNet18	50.2	56.9	54.4	63.9	49.3	54.1	48.7	54.6
MoCo-v1	ResNet18	49.8	55.1	53.1	61.7	47.2	51.9	47.9	53.0
MoCo-v2	ResNet18	55.2	63.6	58.3	68.3	51.8	57.6	52.1	58.7
DenseCL	ResNet18	56.0	64.8	58.8	68.7	52.5	58.2	52.9	60.0
ConCL	ResNet18	58.6	68.1	61.0	71.3	55.1	61.4	54.5	62.6
MIM, i. e., the backbone weights are transferred from masked image modeling.									
SparK	ResNet18	56.0	65.8	57.4	67.6	53.4	59.7	53.0	59.6
MAE	ViT-S/16	57.2	65.3	59.4	68.8	56.0	61.9	55.8	61.8
BeitV2	ViT-S/16	57.1	65.3	59.7	68.1	56.1	63.2	55.8	62.3
EVA	ViT-S/16	57.5	66.2	60.2	69.9	55.2	62.6	55.6	62.0
Ours	ResNet18	**60.8**	**71.1**	**61.7**	**72.1**	**57.1**	**64.4**	**56.4**	**64.3**

Pre-training architecture. We use ResNet18 as the network backbone, and is compared with state-of-the-art approaches that either uses ResNet18 [22] or ViT-S/16 [14] backbones. Note that SSL in pathology images is an emerging area, and the state-of-the-art methods from the natural image domain are benchmarked by [47] or us. Refer to [47] for optimization details of contrastive learning approaches, and we use batch size 256 for benmarking MIM methods.

Baseline. We compare with four method groups: i) training from scratch. Random weight initialization is used for training on downstream tasks; ii) supervised pre-training. We initialize the backbone weights with pre-training on the NCT dataset for tissue classification [26, 47]; iii) contrastive learning. The backbone weights are initialized by using SimCLR [5], BYOL [17], MoCo-v1 [20], MoCo-v2 [6], DenseCL [44], or ConCL [47]; iv) MIM. SparK [43], MAE [19], Beitv2 [36], or EVA [15] pre-trained backbone weights are transferred to downstream tasks.

Downstream task & dataset. Following [47], we explore object detection and instance segmentation of glands, and use the pathology images challenge (GlaS) dataset [41] and Colorectal adenocarcinoma gland (CRAG) dataset [18].

Downstream task architecture. We use Mask RCNN [21] with a backbone network and a feature pyramid network (FPN) head [32] as the detector, and following [47], the fully convolutional network (FCN) head from [21] is used for segmentation. For our method and other convolutional state-of-the-art approaches [6, 17, 20, 22, 43, 44, 47], the Mask RCNN backbone network is set to ResNet18. For the approaches with a ViT backbone [14, 15, 19, 36], ViTDet [31] is used to construct a feature pyramid from the isometric features of the ViT backbone for the Mask RCNN.

Implementation detail. `mmselfsup` [11] and `Detectron2` [45] codebases are used for pre-training each method and fine-tuning them on the downstream datasets. The average results of 5 independent experiments are reported. In pre-training, we follow [47] for data augmentations, and use mask patch size 32×32 [43]. Our method is trained from scratch with an SGD optimizer for 800 epochs. The learning rate is set to 3×10^{-2}, and is scheduled with a cosine annealing strategy [47]. The weight decay and momentum of the optimizer are respectively 1×10^{-4} and 9×10^{-1}. We use batch size 256. A single decoder layer from [19] with hidden dimension 256 is used as our decoder. Our fine-tuning pipelines follow [47] and [31].

4.2. Experimental Result

We use the family of mean average precision (mAP) metric for evaluations, following COCO-style [34, 47], i. e., AP^{bb}/AP^{mk} and AP^{bb}_{75}/$^{mk}_{75}$ (the higher the better) for the detection/segmentation task.

The comparisons with state-of-the-art approaches on the

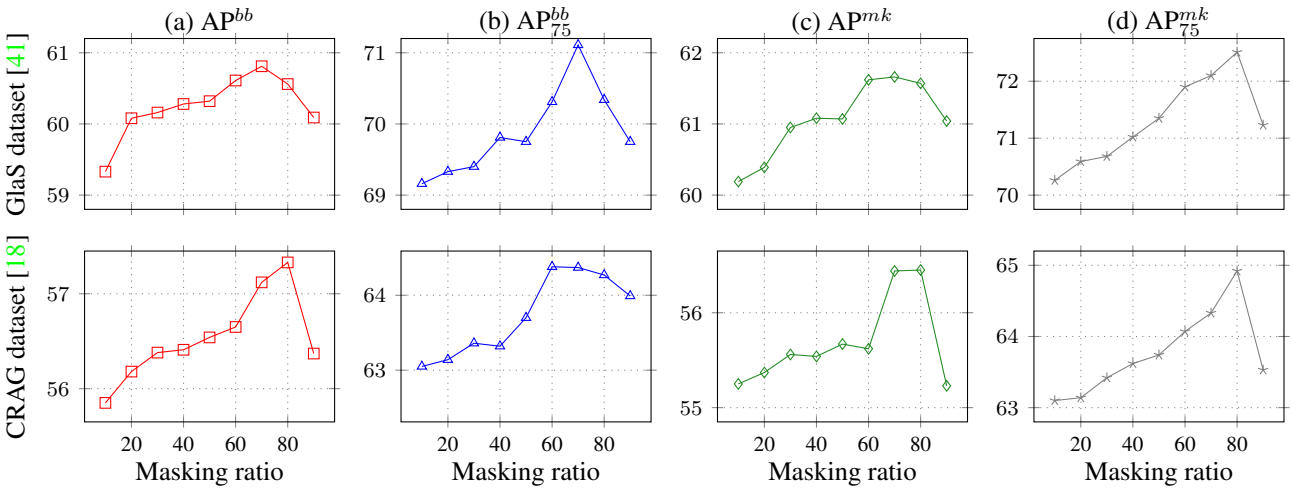


Figure 4. *The object detection and instance segmentation performance of our method with respect to different masking ratios on the GlaS dataset [41] and CRAG dataset [18], corresponding to the first and second rows. We present (a) AP^{bb}, (b) AP^{bb}_{75}, (c) AP^{mk} and (d) AP^{mk}_{75} from the first to last columns repscetively. Overall, the masking ratio 70% finds the best performance. Best viewed in color on the screen.*

GlaS dataset [41] and CRAG dataset [18] are given in Tab. 1. Per the comparisons, our findings are as follows: i) the groups of 'training from scratch' and 'supervised pre-training' obtain similar performance, suggesting that there is barely beneficial knowledge from the tissue classification tasks to the downstream tasks of dense prediction, i. e., object detection and instance segmentation. This further demonstrates the demands of learning discriminate features for the dense prediction tasks in a self-supervised pre-training manner; ii) ConCL and EVA tend to achieve the second-best performance, with a ResNet18 and ViT-S/16 backbone. Though ViT-S/16 almost doubles the number of parameters in ResNet18, their performance stays similar. With the observation, we are motivated to explore the convolutional backbone-based self-supervised learning framework; iii) the past convolutional MIM approach, SparK, is less competitive, compared to the dense contrastive learning approaches [44, 47]. As analyzed in Sec. 1, features of pathology images usually have a large span, and the sparse convolution used by SparK prevents the network from modeling such features, potentially incurring domain shifts between model pre-training and downstream tasks fine-tuning; *iv)* our method consistently finds the best performance across all evaluation metrics on the two downstream task datasets.

4.3. Ablation Study

We ablate our network components and optimization strategy for both pre-training and fine-tuning.

Masking ratio. We study the masking ratio for pre-training our framework. The masking ratio is varied from 10% to 90% with a step size of 10%, and the results are given in Fig. 4. Setting the masking ratio as 70% tends to find the overall best performance. This finding is consistent with [19] that a high masking ratio is hypothesized to learn discriminative features for downstream tasks. It is also worth noting that our method outperforms the past convolutional MIM approaches [43] even with other sub-optimal masking ratios, suggesting robustness of our approach.

Table 2. *Ablation study of model components.*

Method	GlaS Dataset		CRAG Dataset	
	AP^{bb}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{75}
SparK [43]	56.0	65.8	53.4	59.7
Emb. masking	59.9	70.0	56.6	64.0
Emb. invariance	58.7	68.6	56.5	63.6
Ours	**60.8**	**71.1**	**57.1**	**64.4**

Model component. We study our self-supervised learning components. We consider three baselines: i) SparK [43]. This state-of-the-art method can be considered as a baseline for our convolutional MIM framework; ii) Emb. masking. We use our embedding masking strategy for pre-training; iii) Emb. invariance. Our strategy of learning invariant embedding is used in pre-training. Our component consistently improves the downstream task performance. Compared to the second-best baseline, 'Emb. masking', we have 0.9/1.1 and 0.5/0.4 higher AP^{bb}/AP^{bb}_{75} on the GlaS dataset [41] and CRAG dataset [18] respectively.

Fine-tuning schedule. We study the fine-tuning schedules for transferring our pre-trained backbone weights on the downstream tasks (Tab. 3). We follow the same setting as [47], where the standard fine-tuning schedule is set as '1×'. In other settings, the digit before '×' indicates a multiplier for scaling the optimization iterations in the stan-

Table 3. *Performance of our method on the detection and instance segmentation downstream tasks under different fine-tuning schedules. The standard fine-tuning schedule [47] is defined as 1×, and is scaled to 0.5×,2×, 3×, and 5× schedules.*

Schedule	GlaS dataset				CRAG dataset			
	Detection		Segmentation		Detection		Segmentation	
	AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{75}
0.5×	60.0	70.3	60.7	71.0	57.5	65.0	**56.7**	65.2
1×	60.8	71.1	61.7	72.1	57.1	64.4	56.4	64.3
2×	61.0	71.2	61.8	72.2	57.6	**65.2**	56.5	65.4
3×	60.9	**71.3**	**62.1**	**72.6**	**57.7**	64.9	56.5	65.2
5×	**61.1**	70.9	61.9	72.3	**57.7**	65.1	**56.7**	**65.5**

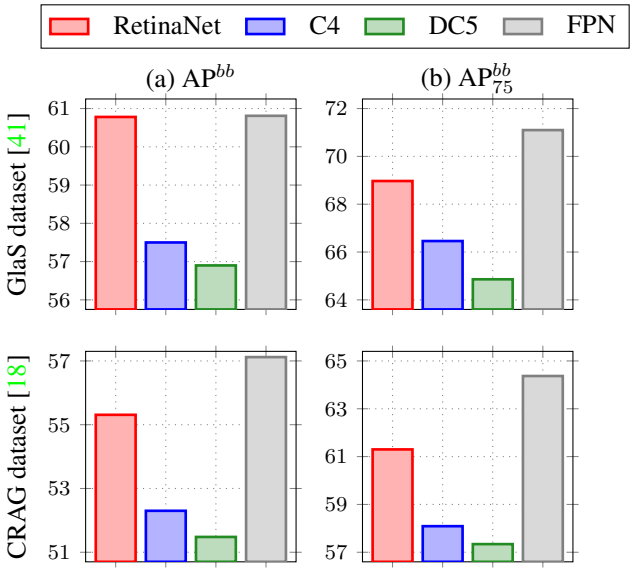


Figure 5. *Ablation of the detector backbone on the GlaS dataset (the first row) and the CRAG dataset (the second row). We use (a) AP^{bb} and (b) AP^{bb}_{75} as the evaluation metrics. We explore the four well-known detectors: RetinaNet [33], RCNN-C4 (dubbed C4) [38], RCNN-DC5 (dubbed DC5) [38], and RCNN-FPN (dubbed FPN) [32, 38]. Best viewed in color on the screen.*

dard fine-tuning schedule. Our approach can converge with 0.5×, 1×, or 2× schedules. Unlike ConCL [47] that benefits most with the 5× schedule (indicated by its Tab. 3), our approach converges faster, showing that our model learns more dense prediction knowledge during pre-training.

Detector architecture. We study the impact of the detector architecture for transferring our pre-trained backbone network in the downstream tasks. The comparisons are given in Fig. 5. We consider four well-known detecor architectures, RetinaNet [33], RetinaNet [33], RCNN-C4 [38], RCNN-DC5 [38], and RCNN-FPN [32, 38]. The RCNN-FPN architecture significantly outperforms other detector architectures on the downstream tasks. Even with the least performed RCNN-DC5 detector, we remain to find 56.9 AP^{bb} and 56.9 AP^{bb}_{75} on the GlaS dataset [41], which outperforms most of the state-of-the-art approaches from Tab. 1.

Table 4. *Ablation study of pre-training epochs.*

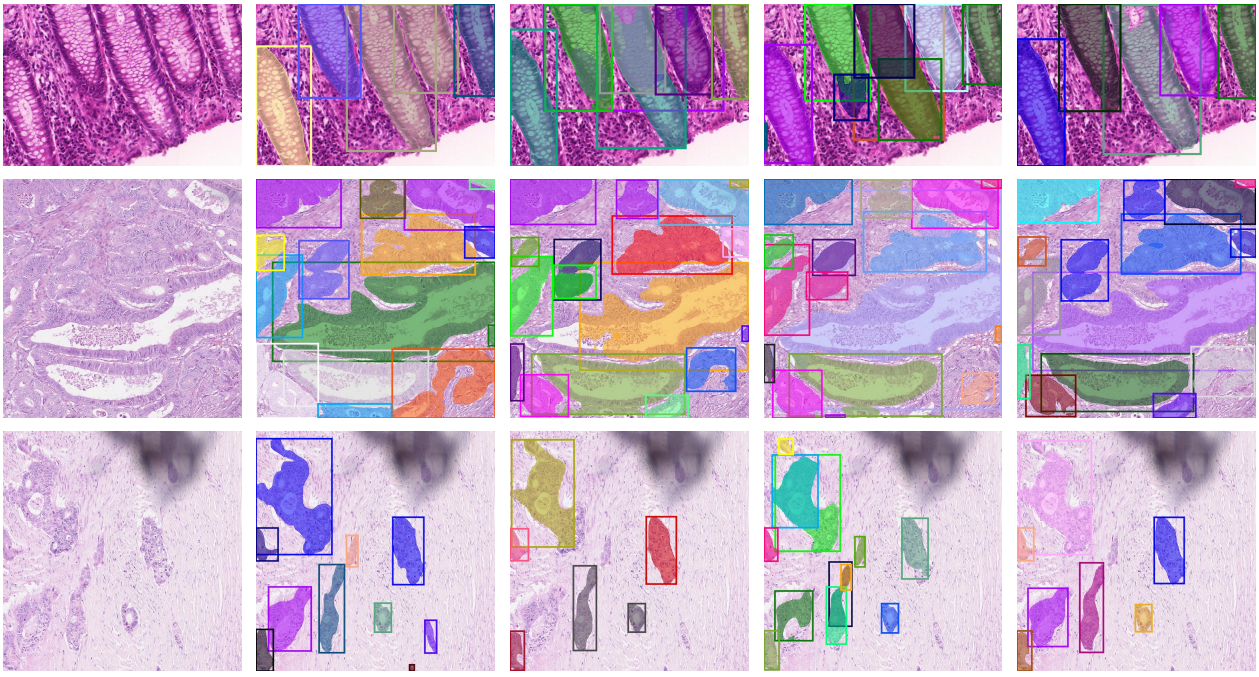
Pre-training Epoch	GlaS Dataset		CRAG Dataset	
	AP^{bb}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{75}
200	58.5	65.4	56.3	63.4
400	59.7	69.1	56.8	63.7
800	**60.8**	71.1	57.1	64.4
1600	**60.8**	**71.2**	**57.4**	**64.6**

Pre-training epoch. We respectively pre-train our methods for 200, 400, 800, and 1600 epochs to study the relations between pre-training epochs and downstream task performance. The results are given in Tab. 4. Consistently, the downstream task performance benefits from increments of the pre-training epochs. However, by pre-training our method for 800 epochs, computation costs and downstream task performance are balanced on both datasets.

Model scalability. We scale our backbone network to ResNet34 and ResNet 50 in Tab. 5, and pre-train them with the same settings. We find that the best performance on the GlaS dataset [41] and CRAG dataset [18] is achieved by using the ResNet34 and ResNet50 backbone respectively. One may note that the ResNet50 backbone has a performance decrease on the GlaS dataset [41]. This can be explained that we use the same fine-tuning pipelines for all scales of model architectures [47]. A large model always requires stronger regularization than a small model when fine-tuning on the same dataset [19, 47]. In this paper, we focus on the study of network pre-training, and do not further investigate the fine-tuning pipelines.

5. Discussion

Detection and segmentation visualization. We visualize the detected bounding box and segmented masks of our method and the two most competitive methods from Tab. 1, i. e., ConCL [47] and EVA [15], in Fig. 6. The results show that our method can capture objects better on both the detection and segmentation tasks, benefiting from our effectively pre-training strategies. For example, in the first row, our method accurately detects the five glands, while ConCL and EVA results in 2 and 3 false positive detection.



(a) Image (b) Ground truth (c) ConCL [47] (d) EVA [15] (e) Ours

Figure 6. *Detection and instance segmentation examples of state-of-the-art approaches and our method. (a) is pathology images. (b) is ground truth bounding boxes and instance segmentation masks. (c) and (d) are predictions of ConCL [47] and EVA [15], the most performing past approaches in our tasks. (e) is predicted by our method. Best viewed in color on the screen.*

Table 5. *Ablation study of model scalability.*

Backbone	GlaS Dataset		CRAG Dataset	
	AP^{bb}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{75}
ResNet18	60.8	71.1	57.1	64.4
ResNet34	**61.5**	**72.2**	58.0	65.3
ResNet50	60.5	69.8	**58.3**	**66.1**

Table 6. *Comparison with foundation model of instance segmentation, SAM [27]. The '-' symbol denotes unavailability.*

SAM Prompt	GlaS Dataset		CRAG Dataset	
	AP^{bb}	AP^{mk}	AP^{bb}	AP^{mk}
Automatic	1.9	1.6	4.4	4.1
GT point	-	11.7	-	21.0
GT bbox	-	58.2	-	57.2
Ours	**60.8**	**61.7**	**57.1**	**64.4**

Comparison with SAM. We compare with the recently released foundation model in our task, i. e., SAM [27] that is trained with a ViT-H backbone on billions of instance masks. Three types of prompts for SAM are explored: i) Automatic. We use the official automatic point prompts generator from SAM for instance segmentation. Bounding box predictions are inferred from the segmentation mask; ii) GT point. We supply the center point of each ground truth bounding box to SAM; iii) GT bbox. The ground truth bounding box is used to prompt SAM for the instance segmentation task. The 'GT bbox' setting can be considered as the upper-bound performance of SAM. Consistently, our method finds a better performance than SAM.

6. Conclusion and Broader Impact

In this paper, we study a self-supervised convolutional MIM framework for facilitating dense prediction tasks on pathology images. Observing that the pathology images contain features of large spatial span, different affine shapes, and diverse color, our key insights are using `[MASK]` tokens after the stem layer of the work for boosting information propagation through masked regions during the encoding stage, and encouraging the network to learn invariant embedding through constraining the reconstruction targets during the decoding stage. With extensive experiments, our method demonstrates state-of-the-art transfer learning performance on standard benchmark datasets.

Broader impact. Our method can be potentially applied to assist in modern health care. We hope this paper will draw more attention to SSL on pathology images.

Acknowledgment. Liyuan Pan's work was supported in part by the Beijing Institute of Technology Research Fund Program for Young Scholars and National Natural Science Foundation of China 62302045.

References

[1] Christian Abbet, Linda Studer, Inti Zlobec, and Jean-Philippe Thiran. Toward automatic tumor-stroma ratio assessment for survival analysis in colorectal cancer. In *Medical Imaging with Deep Learning*, 2022. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 2, 3, 5

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE, 2021. 3

[8] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 2, 3

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[10] Ozan Ciga, Anne L. Martel, and Tony Xu. Self supervised contrastive learning for digital histopathology. *CoRR*, abs/2011.13971, 2020. 3

[11] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021. 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 3, 4, 5

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: exploring the limits of masked visual representation learning at scale. *CoRR*, abs/2211.07636, 2022. 5, 7, 8

[16] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10032–10042. IEEE, 2021. 3

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 5

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. 06 2020. 5, 6, 7

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 3, 4, 5, 6, 7

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

2020, Seattle, WA, USA, June 13-19, 2020, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. 2, 5

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1, 2, 4, 5

[23] Md. Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, Taneli Mattila, and Tapio Seppänen. Whole slide image registration via multi-stained feature matching. *Comput. Biol. Medicine*, 144:105301, 2022. 2

[24] Peng Jiang and Srikanth Saripalli. Contrastive learning of features between images and lidar. In *18th IEEE International Conference on Automation Science and Engineering, CASE 2022, Mexico City, Mexico, August 20-24, 2022*, pages 411–417. IEEE, 2022. 3

[25] Hongtao Kang, Die Luo, Weihua Feng, Li Chen, Junbo Hu, Shaoqun Zeng, Tingwei Quan, and Xiuli Liu. Stainnet: a fast and robust stain normalization network. *CoRR*, abs/2012.12535, 2020. 2

[26] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Medical Image Analysis*, 2018. 2, 4, 5

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *CoRR*, abs/2304.02643, 2023. 2, 8

[28] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir M. Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Trans. Medical Imaging*, 40(10):2845–2856, 2021. 3

[29] Dr Kumar. Survey of machine learning applications of convolutional neural networks to medical image analysis. *International Journal for Research in Applied Science and Engineering Technology*, 9:1186–1196, 11 2021. 1

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[31] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 280–296. Springer, 2022. 5

[32] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017. 5, 7

[33] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 7

[34] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 1, 5

[35] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *CoRR*, abs/2004.09666, 2020. 1

[36] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *CoRR*, abs/2208.06366, 2022. 3, 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 3

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 7

[39] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1144–1153. Computer Vision Foundation / IEEE, 2021. 3

[40] Yash Sharma, Yi Zhu, Chris Russell, and Thomas Brox. Pixel-level correspondence for self-supervised learning from video. *CoRR*, abs/2207.03866, 2022. 3

[41] Korsuk Sirinukunwattana, Josien Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Guo, Li Yang Wang, Bogdan Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David Snead, and Nasir Rajpoot. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35, 03 2016. 1, 5, 6, 7

[42] Chetan Srinidhi, Ozan Ciga, and Anne Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 01 2021. 1

[43] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. *CoRR*, abs/2301.03580, 2023. 1, 2, 3, 4, 5, 6

[44] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3024–3033. Computer Vision Foundation / IEEE, 2021. 3, 5, 6

[45] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[46] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[47] Jiawei Yang, Hanbo Chen, Yuan Liang, Junzhou Huang, Lei He, and Jianhua Yao. Concl: Concept contrastive learning for dense prediction pre-training in pathology images. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXI*, volume 13681 of *Lecture Notes in Computer Science*, pages 523–539. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[48] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. *arXiv preprint arXiv:2301.01928*, 2023. 3

[49] Yan Yang, Liyuan Pan, Liu liu, and Eric A Stone. Isg: I can see your gene expression. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1

[50] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3