Multiple Instance Captioning: Learning Representations from Histopathology Textbooks and Articles

Jevgenij Gamper
University of Warwick, UK
j.gamper@warwick.ac.uk

Nasir Rajpoot
University of Warwick, UK
n.m.rajpoot@warwick.ac.uk

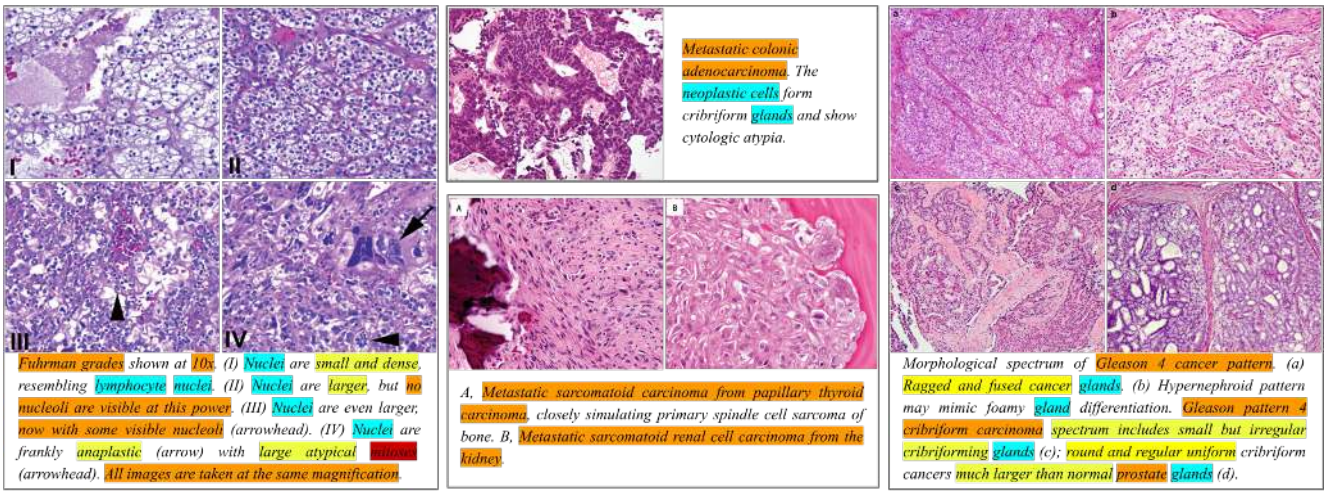


Figure 1: Four samples from ARCH, a multiple instance captioning computational pathology dataset. Samples on the left and right each consist of four image instances with a single caption; top-middle shows an image-caption pair while bottom-middle contains two image instances with a single caption. Labeled in color are examples of common tasks within computational pathology: diagnostic (orange); detection & classification (cyan); descriptive (yellow); special cell detection (red).

Abstract

We present ARCH, a computational pathology (CP) multiple instance captioning dataset to facilitate dense supervision of CP tasks. Existing CP datasets focus on narrow tasks; ARCH on the other hand contains dense diagnostic and morphological descriptions for a range of stains, tissue types and pathologies. Using intrinsic dimensionality estimation, we show that ARCH is the only CP dataset to (ARCH-)rival its computer vision analog MS-COCO Captions. We conjecture that an encoder pre-trained on dense image captions learns transferable representations for most CP tasks. We support the conjecture with evidence that ARCH representation transfers to a variety of pathology sub-tasks better than ImageNet features or representations obtained via self-supervised or multi-task learning on pathology images alone. We release our best model and invite other researchers to test it on their CP tasks.

1. Introduction

The success of an intelligent system depends on having the right representation of data. Computer vision community has gradually moved from engineering features to letting the deep neural networks learn data representations, given a differentiable task objective [1, 18].

Computational pathology (CP), a sub-field of medical imaging that entails quantitative profiling of spatial patterns in multi-gigapixel whole-slide images (WSIs) of patient tissue slides, has followed these developments closely [10]. Deep Learning (DL) has been applied to detecting cancerous regions [34], classifying tissue sub-types [28], identifying diagnostically relevant structures such as glands [21], nuclei [17], vessels and nerves [15], quantifying spatial patterns of tumor infiltrating lymphocytes [53] and histology image retrieval [23]. More recently, DL been used to learn representations for challenging tasks of predicting genetic sub-types [16, 31].

With recent advances in weak and unsupervised learning,

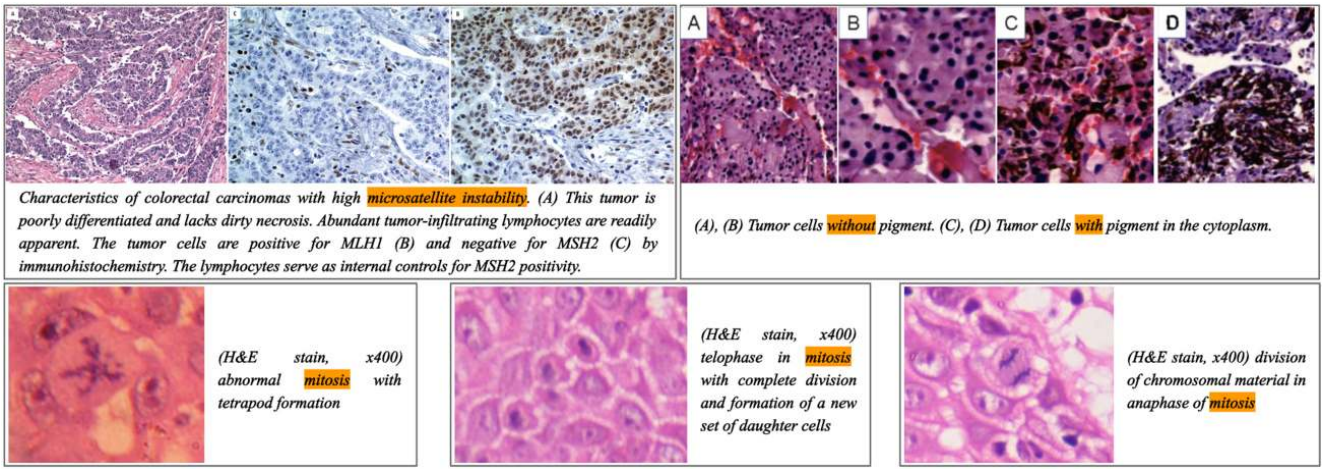


Figure 2: Five samples from ARCH. *Top left*: Caption describes morphological characteristics of Microsatellite Instability (MSI) in colon cancer; *Top right*: Caption provides contrastive supervision where cells in first two images have to be contrasted against pigmented cells in the last two images; *Bottom row*: Captions provide morphological description of cell-division or *mitosis* at different stages.

it is becoming widely acceptable to design computer vision systems from pre-trained representations [58, 19, 24, 29, 43, 20]. CP community has followed suit, by relying less on patch-level annotations and instead adopting weakly supervised methods such as multiple instance learning (MIL) [6, 45, 26] or neural image compression (NIC) [57] to perform inference directly from WSI-level data. However, most of the above methods rely on encoders pre-trained on ImageNet for feature extraction.

These developments create a growing demand for pre-trained pathology image representations. In this paper, we present ARCH – a new dataset of dense image captions mined from clinical and academic pathology textbooks and articles – and demonstrate the universality of pathology image features obtained by pre-training on ARCH.

We conjecture that *the global minimum of the multiple instance captioning objective on a dataset like ARCH is the global minimum of the majority of computational pathology sub-tasks commonly solved and published to date*. We base this conjecture on several observations made while putting together this dataset and on experimental results reported in this work. In Figure 1, correctly captioning a bag of four image instances (left) requires an algorithm to: identify nuclei as well as their category (cyan) [17], describe nuclear characteristics (yellow) [40], detect mitotic figures (red) [9, 61] and even understand the relationship between image magnification and diagnosis (orange). Besides identifying cells and glands (all examples, cyan) [15, 21], an algorithm has to learn to describe an image as being metastatic tumour originating in colon (middle-top) or a metastatic sarcoma from papillary thyroid (middle-bottom). Only recently, Lu *et al*. [46] proposed a method for addressing the challenging task of identifying metastatic cancers and tumor origins. In Figure 1 on the right, an algorithm has to identify the bag of four images as prostate and assign it a Gleason grade 4 [48, 3, 5]. The ARCH multiple instance captioning dataset stands out by offering implicit supervision of contrastive learning – where instances within the same bag have to be contrasted in order to generate the caption (Figure 2 top-right). ARCH includes samples that provide explicit supervision on genetic characteristics expressed through tissue morphology (Figure 2 top-left), a task that has only recently been explored [33, 31, 16]. One may also find close up descriptions of mitosis at various stages (Figure 2 bottom row) [52]. We provide additional samples from ARCH in the Supplementary Material. We also present attention plots (see Figure 3) as additional evidence to support our conjecture and demonstrate examples at low and medium resolutions that are widely available in ARCH.

While ARCH includes unique CP tasks as sub-tasks to image captioning, we believe it is the relationship between the sub-tasks described in captions that provides unique supervision during pre-training and results in more general CP image features. In Section 3.2, we compare the intrinsic dimensionality of CP datasets to their analogs in computer vision. CP datasets appear to have lower intrinsic dimensionality than CIFAR100 and CityScapes in their respective tasks of image classification and segmentation. ARCH, on the other hand, matches the complexity of its computer vision counterpart, the MS-COCO Captions dataset. Recently, Desai *et al*. [12] showed that pre-training on MS-COCO captions is superior to ImageNet. This might provide a hint to why ARCH pre-training is advantageous to multi-task training. Individual CP tasks provide narrow su-

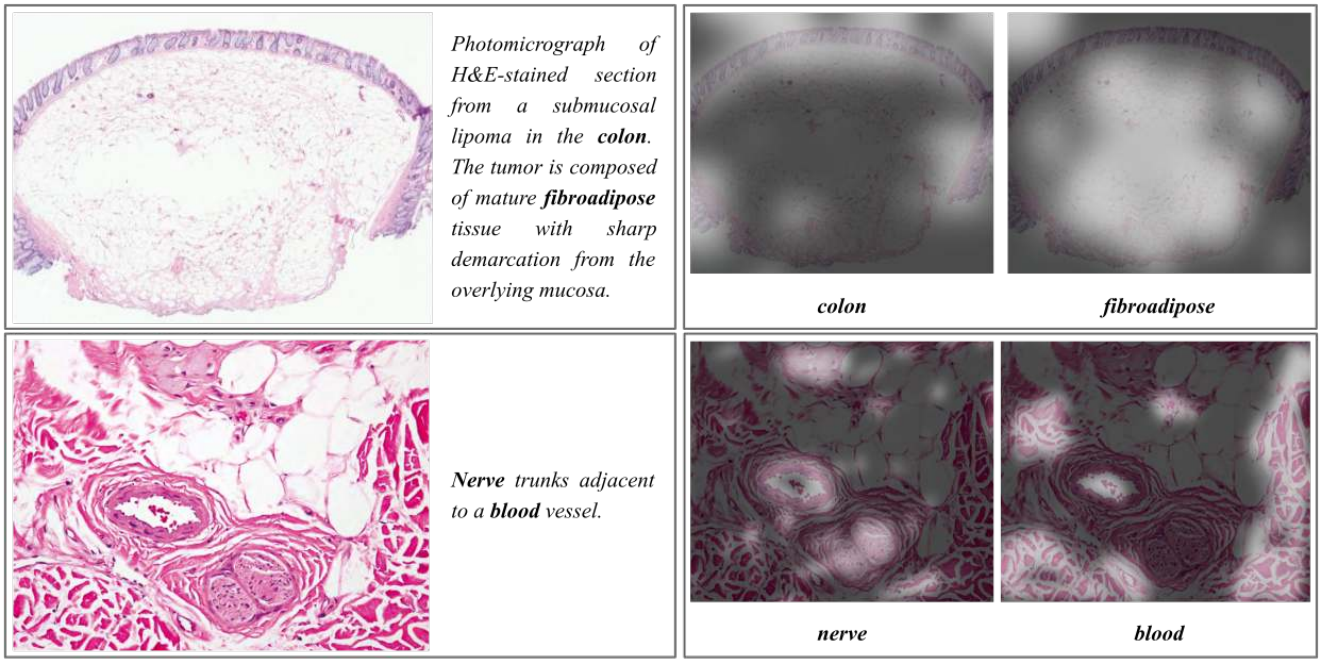


Figure 3: *Left*: samples of tissue at low and higher power view with their corresponding captions; *Right*: Model generated attention masks the corresponding words highlighted in bold. Attention masks were generated according to the methods described in Desai *et al*. [12], see Section 4.2 for details.

pervision and even when combined together in a multi-task setting, the supervision signal does not explicitly model the relationship between the tasks.

Finally, besides offering a new computer vision task of multiple instance captioning, ARCH presents a previously unexplored path towards providing information rich ground truth for medical imaging, a field known for its annotation scarcity.

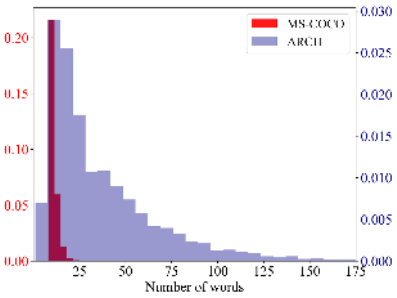
2. Related Work

Desai *et al*. [12] showed that models pre-trained on MS-COCO Captions efficiently transfer to classification, and detection tasks. MS-COCO contains only 118k images, much smaller in number than ImageNet, yet the authors of [12] have shown that learning from textual annotations is more data-efficient than from classification labels. They emphasized superior results from using more images with single captions, rather than multiple captions per image; this observation is particularly important for ARCH, where we only have a single caption per bag of images. In this work, besides offering a new dataset for multiple instance captioning in CP, we show that pre-training on single image captions is indeed more data-efficient and induces superior inductive biases for data representation as compared to multi-task or self supervised pre-training.

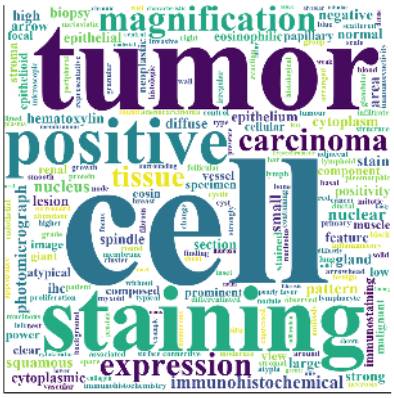
We use strict measures to compare pre-trained representations to support our conjecture. The standard methodology has been to extract features from the last layer of the trained encoder and evaluate them with a linear classifier [14, 55, 8], to which Resnick *et al*. [51] referred to as *linear probing* and showed it to be insufficient if downstream tasks are highly non-linear. In this work we benchmark the performance of an encoder pre-trained on ARCH against encoders pre-trained on ImageNet, via self- or multi-task supervision evaluated with the encoder that produces the best performance, while our model can only use linear encoder.

Transferring models from ImageNet remained to be the dominant method in practice until recently, however pre-training on textual annotations and self-supervised learning are closing the performance gap. Kornblith *et al*. [37] offered an empirical study on the transferability properties of models trained on ImageNet. Raghu *et al*. [50] further studied ImageNet models transferability for medical imaging datasets and showed that the primary benefit of transfer has been for the initialisation of large model weights rather than feature transfer and that the same result could be obtained with far smaller models trained from scratch.

Meanwhile, the CP community has pursued two options to pre-train features that generalise across datasets and tasks to tackle the problem of the scarcity of ground-truth data: Multi-Task learning (MTL) and self-supervised (SS) learning [44]. Tellez *et al*. [56] recently extended NIC via multi-task pre-training on four tasks. Mormont *et al*. [47]



(a)



(b)

Bag size	1	2	3
# Bags	9,772	1,292	461
Total Bags		11,816	
Bag size	4	5	6
# Bags	173	44	36
Total Images		15,164	
Bag Size	7	8	9
# Bags	14	17	7

(c)

Figure 4: (a) Histogram of caption lengths showing that MS COCO captions are significantly shorter as compared to captions in ARCH; (b) Word cloud of 1,000 most frequent words in ARCH captions; (c) Tables of bag sizes – i.e., number of images per single caption in ARCH bags.

provided an in-depth analysis of feature generalization of multi-task pre-training on 22 image classification tasks (11 of the used datasets are not publicly available). In this work, besides introducing a unique pre-training dataset ARCH, we use only publicly available datasets for evaluation.

Even though our goal is to obtain general image features from pre-training on image captions, not image captioning in itself, it is important to point out the work by Zhang *et al*. [63], who were the first to propose a pathology image captioning model and dataset. Yet the dataset used for training was small - 32 patient images that are not publicly available, limited to single tissue type (bladder) and a single stain type (H&E). Our pathology image captions dataset is substantially larger, and includes a whole range of tissues, diagnoses and stain types.

3. Method

3.1. ARCH Construction

We used PubMed medical articles database and pathology textbooks to construct ARCH. We selected all PubMed journals according to keywords (*pathology*, *histochemistry*, *histology*, *histopathology*), which resulted in 12,676 journal articles on clinical and research pathology as of 2019. Using *pubmed parser*[1], we extracted a total of 25,028 figures and their corresponding captions. We then manually selected 8,617 figure-caption pairs that contained at least one histology or imunohistochemistry (IHC) image and we saved text in caption that only related to the histology image. Individual images were then extracted from figures to create multiple instance bags with their respective captions that have been checked for formatting errors. When selecting images, we made sure that these did not include excessive text, marks and were of reasonable quality.

Using *PDF-Figures 2.0*[2], we extracted figures and their captions from 10 textbooks, leading to 3,199 figure-captions pairs. We followed the above steps for constructing multiple instance bag and caption pairs, verifying the quality of the images as well as checking extracted captions for formatting errors.

ARCH contains 11,816 bags and 15,164 images in total. Figure 4c shows a more detailed breakdown by the number of bags according to the number of images within the bag, with the smallest bag size being 1 (9,772 samples) and the largest bag size being 9 for which we have only 7 samples.

3.2. ARCH Intrinsic Dimensionality

Li *et al*. [41] demonstrated how a number of parameters in a randomly oriented subspace of a fixed neural network architecture can be used as a rough gauge of the difficulty of the problem. One slowly increases the dimension of this subspace, monitors at which dimension size the solutions first appear, and defines that to be the *intrinsic dimensionality*. We follow this methodology to compare the relative complexity of CP datasets and their computer vision analogs, as shown in Figure 6. [3]. The three tasks are image classification, image segmentation and image captioning, with their respective evaluation metrics of accuracy, panoptic quality [35] and CIDEr [60] as per Desai *et al*. [12]. For each of the tasks, as well as for all of our experiments in the paper, we used ResNet-18 [22] as an encoder with its parameters wrapped in Fast-food transform [39] for obtaining random sub-spaces required for intrinsic dimensionality estimation. All encoders were randomly initialised.

[1]https://github.com/titipata/pubmed_parser

[2]https://github.com/allenai/pdffigures2

[3]We used a publicly available PyTorch implementation of intrinsic dimensionality estimation github.com/jgamper/intrinsic-dimensionality

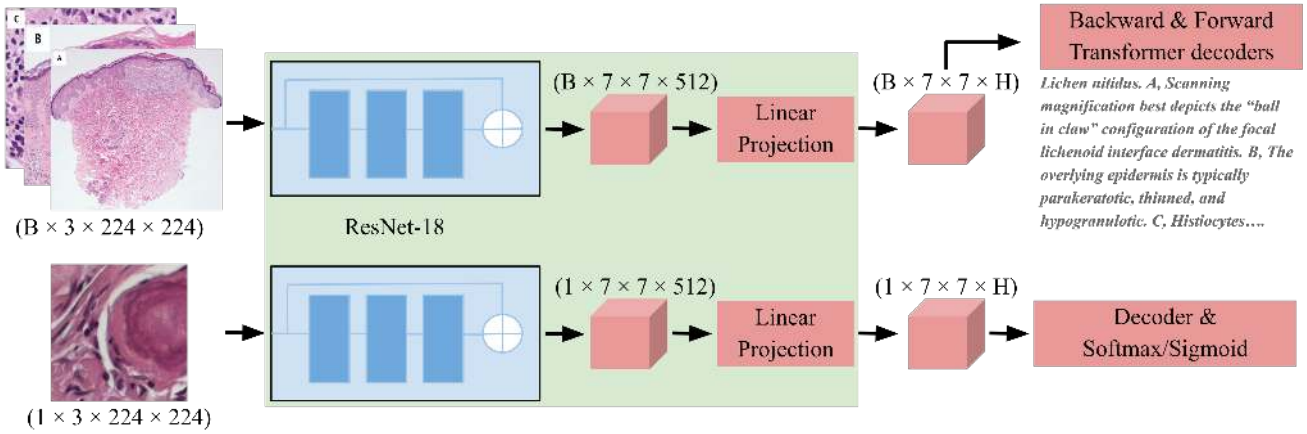


Figure 5: Schema of neural networks used in experiments, all use ResNet-18 encoder. *Top row*: ARCH pre-training architecture. Within a bag all images are encoded, these features are then projected by a linear layer with non-linearity and fed to backward and forward transformers as per Desai *et al*. [12]; *Bottom row*: for image classification, after encoding the features are also projected by a linear layer with non-linearity and fed to sigmoid or softmax depending on the number of labels. In green, we show model components that are shared during multi-task training. For both models, we modify ResNet-18 to have the Batch-norm layer as an input layer.

3.3. ImageNet baseline

To benchmark representations we first need to obtain a competitive baseline. It is a general practice to use penultimate layer of a model trained on ImageNet for feature extraction, however some [45] have found shallow layers to be more useful in CP tasks. In addition, *linear probing* (LP), a method that uses a linear classifier to evaluate the quality of the representations may not be applicable. CP images have no reason to be linearly separable in ImageNet space. We thus search for optimal network depth for feature extraction i.e. the number of hidden layers in the decoder and dropout regularisation that optimally transfers ImageNet features for a given classification task. To achieve that, while keeping weights frozen of ResNet-18 pre-trained on ImageNet, we grid-searched for optimal decoder regularisation across 3 decoder depths (linear, 1, and 3 hidden layers) and 5 encoder depths. See Figure 7 for example results.

In Table 1, a linear decoder on top of penultimate layer of ImageNet trained Resnet-18 is referred to as I-L; a linear decoder with an ImageNet ResNet-18 layer selected through hyper-parameter tunning is referred to as I-L-O; finally, I-B refers to the performance of the best ImageNet ResNet-18 layer followed by the decoder with depth and regularisation that lead to the best performance.

3.4. Self-supervised learning baseline

Lu *et al*. [44] proposed contrastive predictive coding (CPC) as a self-supervised objective to pre-train a feature extractor before training a complete multiple instance learning model on whole slide images. We thus include this as an additional baseline for comparison labelled as SS. The training objective remains the same as in Lu *et al*. [44], except that we use ResNet-18 instead of ResNet-50 as a feature encoder.

3.5. Multiple instance captioning

To pre-train a Multiple Instance Captioning (MIC) model on ARCH, we follow the setup of Desai *et al*. [12]. We show the overall model scheme in Figure 5-top, where the only difference with that of Desai *et al*. [12] is that of working with bags of image instances instead of a single instance. We extend their code-base[4] to work in a multiple instance and/or multi-task setting. Given a dataset of bags of images with captions, our goal is to learn visual representations of images that can be transferred to downstream visual recognition tasks.

A feature encoder (ResNet-18 in our case) is used to process every image within a bag, leading to a representation of $\mathrm{B} \times 7 \times 7 \times 512$ where B is the size of the bag. Fully connected layer follows encoding features to $\mathrm{B} \times 7 \times 7 \times \mathrm{H}$, and fed into forward and backward transformers that perform forward and backward captioning respectively. All of the model components are randomly initialized, and jointly trained to maximize the log-likelihood of the correct caption tokens $C = \{c_1, ..., c_T\}$:

[4]Desai *et al*. [12] code available at github.com/kdexd/virtex. Mormont *et al*. [47] code used for multi-task learning available at github.com/waliens/multitask-dipath.

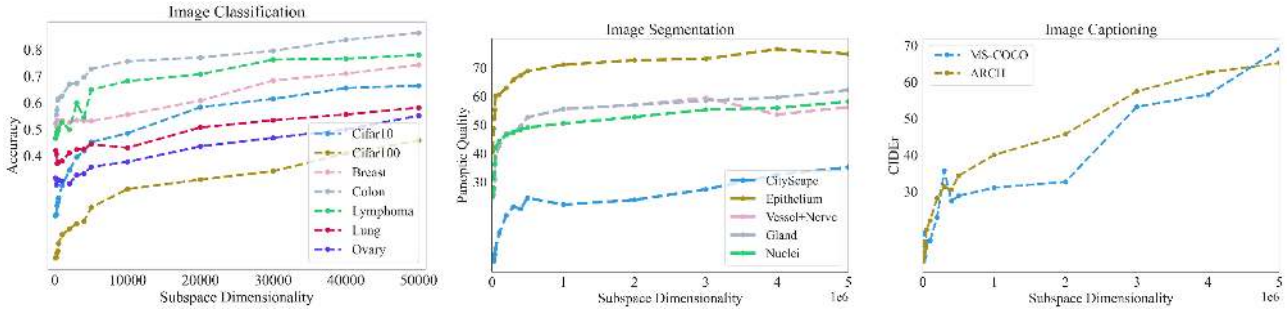


Figure 6: Relative complexity of computer vision and CP datasets for the tasks of image classification (*left*), segmentation (*middle*) and captioning (*right).*

$$\mathcal{L}(\theta, \phi) = \sum_{t=1}^{T} \log(c_t | c_1, ..., c_{t-1}; \theta, \phi_f) \quad (1)$$

$$+ \sum_{t=1}^{T} \log(c_t | c_{t+1}, ..., c_T; \theta, \phi_b) \quad (2)$$

where θ, ϕ_f, ϕ_b are the parameters of the feature extractor and linear layer, forward, and backward models respectively. Tokens are indexed by t. Backward and forward transformer heads are used exclusively in order to backpropogate the supervision signal from the ground truth captions to the feature encoder - again, the goal here is not to learn the best image captioning model.

The performance of the model trained on ARCH is referred to as MIC in Figures 6, 7 and Table 1. For all MIC model training experiments, we set the hyper-paremeters, tokenization and training details according to Desai *et al.* [12] and their publicly available code, with a few exceptions as follows: H is set to 512, which also determines the width of each of the transformer layers and the number of attention heads; the batch size is set to 32 images or less irrespective of the bag sizes due to computational restraints, which are pre-computed before every epoch after re-shuffling the dataset indices. Finally, for the ease of training, we switched to a ADAM optimizer with a default learning rate of 1e-3 and an early stopping set to a patience of a held-out set validation loss of 10.

3.6. Multi-task learning

We also experiment with adding MIC as part of the overall multi-task training with other datasets at hand in the same vein as Mormont *et al.* [47], the idea being that linear decoders may enforce linear separability of the learnt features.

In the MTL setup, the ResNet-18 encoder is shared among the transformer decoders for learning from the caption data and linear decoders for the 12 tasks specified in Table 1. We refer to this model as MTL+MIC, however we also report results of multi-task training without MIC objective as an additional baseline. When training the MTL+MIC model, we had to resort to gradient accumulation to be able to use all datasets at every training iteration.

For all of our MIC, MTL and MTL+MIC models, we employ standard data augmentations using imaug: custom random crop function such that letters are not cropped out); resize and re-scale; color jitter (brightness, contrast saturation and hue); Gaussian and/or salt and pepper noise, and JPEG compression artifacts.

4. Experimental Results

The goal of our experiments section is to show that pre-training a ResNet-18 encoder on ARCH captions allows the encoder to learn transferable features; it is not our goal to learn an image captioning model. As such, we aim to demonstrate the potential of yet unexplored source of supervision – medical textbooks and articles, and their combination with publicly available labelled data.

4.1. Setup

To obtain results in Table 1 and Figure 7 for SS, MTL and MTL+MIC models, we follow the experimental setup described in Mormont *et al.* [47] – i.e., leaving one dataset out. For instance, to test the performance of SS model on the breast cancer classification dataset (Table 1 dataset #1), we exclude that dataset from training and pre-train the SS model on the remaining 11 datasets. The same would apply to MTL and MTL+MIC models, except that we would never exclude ARCH from training in the latter. For ImageNet baselines, we only train on the specific dataset that we are interested in obtaining performance for.

4.2. Results and Discussion

As expected, features for CP images derived from ResNet-18 pre-trained on ImageNet are not linearly separable, as shown by the result for I-L column in Table 1. It is

Num.	Dataset (# classes)	Tissue/Task	I-L	I-L-O	I-B	SS	MTL	MIC	MTL+MIC
1	Litjens *et al*. [42] (2)	Breast	50.9	66.1	72.3	73.2	**91.8**	90.5	90.0
2	Kather *et al*. [32] (9)	Colon	60.6	65.7	79.9	77.4	80.5	81.7	**83.9**
3	Alsubaie *et al*. [2] (6)	Lung	43.4	52.0	60.4	73.4	78.6	82.4	**87.0**
4	Janowczyk *et al*. [27] (3)	Lymphoma	18.6	23.9	26.4	45.2	44.2	42.1	**47.3**
5	Qureshi *et al*. [49] (4)	Meningioma	20.6	36.7	51.1	56.6	70.0	73.9	**84.8**
6	Shaban *et al*. [54] (3)	Head & Neck	33.0	39.1	66.0	68.3	70.9	73.2	**75.2**
7	Kobel *et al*. [36] (5)	Ovary	22.1	31.4	67.6	65.1	72.5	74.7	**77.0**
8	Kather *et al*. [33] (2)	MSI/MSS	61.4	61.4	62.3	64.5	64.0	64.5	**68.1**
9	Hosseini *et al*. [25] (22)	Multi-label	56.9	75.0	**85**.6	77.5	75.8	83.7	85.4
10	Arvaniti *et al*. [3] (5)	Gleason Scoring	18.1	35.3	61.5	68.1	67.2	75.2	**79.4**
11	Veta *et al*. [61]+[27, 52] (2)	Nuclear Mitosis	58.7	59.0	63.4	57.5	68.5	**70.4**	69.3
12	Roux *et al*. [52] (2)	Nuclear Atypia	49.3	52.6	67.3	66.1	69.4	75.1	**80.2**

Table 1: Accuracies for the classification datasets obtained with: I-L, penultimate ImageNet trained ResNet-18 layer and linear decoder; I-L-O best ImageNet trained ResNet-18 layer and linear decoder; I-B best ImageNet trained ResNet-18 layer and optimal decoder depth; SS, a self-supervised learning model described in Section 3.4; MTL, penultimate multi-task trained ResNet-18 layer and linear decoder; MIC, penultimate ARCH trained ResNet-18 layer and linear decoder; MTL+MIC, penultimate ARCH and multi-task trained ResNet-18 layer and linear decoder. Note that dataset #9 is a multi-label classification problem, for which we report average accuracy per binary label to match the original work.

worth noting, however, that there is a considerable amount of variability in the performance of ImageNet features depending on the ResNet-18 block chosen for feature extraction and the number of hidden layers in the decoder. For example, for the colon tissue sub-typing (dataset 2, Table 1), the accuracy can range from 39% to 79% (see Figure 7), and yet remain almost the same across all combinations for the task like MSI/MSS classification. In our experiments, we did not see a clear pattern for an optimal decoder depth and encoder block combination for ResNet-18 pre-trained on ImageNet. This observation implies that although feature encoders pre-trained on ImageNet have achieved good performance in MIL tasks in the literature [45], every new application of MIL requires searching for an optimal combination of feature extractor depth, decoder depth and MIL aggregator. It is also not clear if the performance achieved in MIL by ImageNet models is due to the sufficient discriminative features or due to the MIL aggregators exploiting higher order information at the WSI level.

Results for SS learning in Table 1 show that although SS learning was able to match the performance of the best ImageNet encoder and decoder combination, it was not able to match the performance of the MTL, MIC or MTL+MIC models. It is perhaps worth bearing in mind that SS models are known in practice to be hyper-parameter sensitive and usually need dataset specific design. Therefore the results we report may under-estimate the SS model performance. We would argue that supervised signal should be used when possible.

Finally, following methods described in Desai *et al*. [12], we demonstrate attention plots for selected tokens in Figure 3. In a colon tissue sample, the model appears to be focusing more on the main constituents of the colon tissue. For adipose (fatty) tissue sub-type, the model clearly focused on the middle, adipose rich area of the tissue slide. It is precisely these qualities that we believe provide some supporting evidence to our conjecture and explain the performance of the feature transfer from MIC or MTL+MIC to new tasks. Likewise, for functional qualities of a tissue image such as nerves, our model appears to identify correct areas where these are present. This allows both MIC and MTL+MIC models to perform well on the Atlas dataset [25] whose labels identify functional traits of tissue in the image (dataset #9 in Table 1).

The results above show the potential of using all available supervisory signals for pre-training general CP feature encoders. We also investigated the relative complexity of CP datasets and their computer vision counterparts within respective tasks. Intuitively, one can expect the intrinsic dimensionality of a dataset and a particular task to correlate well with the ability to learn general features from the dataset. Using intrinsic dimensionality estimation method described in Section 3.2, we compared datasets #1, 2, 3, 4, 5, 6 and 7 with CIFAR100 and CIFAR10 [38] for classification; nuclei [17], epithelium [27] nerve and vessel [15] and gland [4] segmentation datasets with CityScape [11]; and ARCH with MS-COCO for image captioning. We use the inverse of the area under the curve for a given dataset as a proxy to the relative intrinsic dimensionality of the dataset, shown in Figure 6. For both classification and segmentation tasks, we find that CP datasets are relatively less complex than their computer vision counterparts, particularly in the case of segmentation. ARCH, on the other hand, closely matches the complexity of MS-COCO – we suspect that

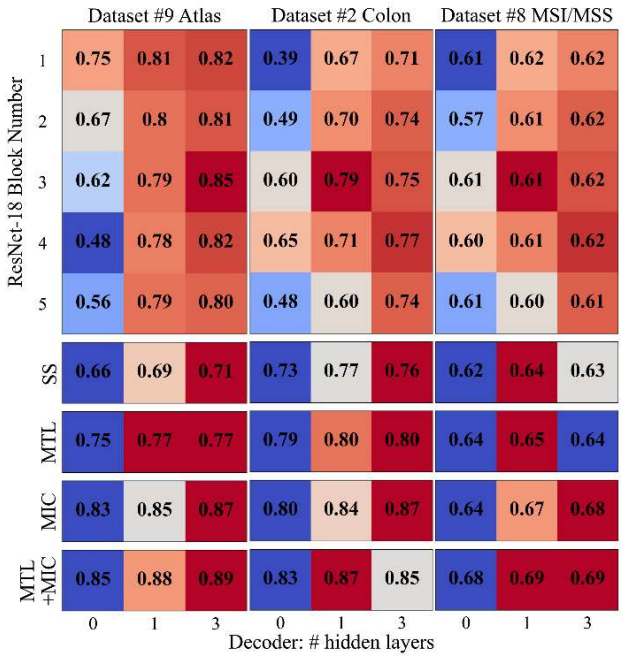


Figure 7: A study of pre-trained feature performance (y-axis) vs number of hidden layers in the decoder (x-axis). First five rows correspond to performance of features extracted from 5 residual blocks at a different depth in ResNet-18 trained on ImageNet. The remaining models correspond to: SS - self-supervised described in Section 3.4; MTL - multi-task learning model described in Section 3.6; MTL+MIC, a multi-task model trained along with ARCH dataset. All features are evaluated with a decoder with 0 (linear), 1 and 3 hidden layers with regularisation optimisation via grid-search. See Supplementary Material for all 12 datasets.

although ARCH is relatively smaller than MS-COCO, the size of the captions (see Figure 4a) and the density of the information within them along with multiple instance nature of the dataset makes the learning process harder.

5. Limitations & Future Directions

Experiments are limited to patch-level performance. While we demonstrate the advantage of pre-training MIC or MTL+MIC models for transferring to completely unseen tasks, the real testing ground would be to evaluate the trained feature extractors on MIL or NIC tasks at the WSI level. Nonetheless, we are confident that there is a significant advantage in MTL+MIC pre-training as compared to transfering features from ImageNet in: ease of use - no need to optimise over network depth - and likely higher performance. While most of the datasets used in large-scale MIL studies are not publicly available, by making ARCH and used datasets publicly available in a single, easy to use repository we hope that the community will make the best use of the trained models.

Dataset construction leads to under-utilisation of transformer based architecture. This work is a result of a laborious effort to extract and clean figures and their captions. While ARCH captions are dense in information, a large majority of the information in clinical papers and textbooks is discarded when constructing ARCH. Dosovitskiy *et al*. [13] have demonstrated that one could employ transformer based models only for image recognition, this opens the path to efficiently learn from semi-structured data such as clinical papers and textbooks with far less effort and our work clearly demonstrates the potential of such approach.

Explosion in hyper-parameter settings. One of the reasons for obtaining an encoder that provides general pathology image features is to avoid grid-search over the ImageNet encoder layer vs decoder depth combinations frequently seen in the literature, and consequently achieve wider adoption in practice. However, MTL, MIC or even SS like models are highly sensitive to hyper-parameters. On the one hand, this opens a call for stricter evaluation protocols; on the other, it is an opportunity for a unified pre-training procedure for CP that could utilise labelled patch-level, pixel-level and ARCH like dataset, as well as weakly labelled datasets such as TCGA [62] and GTEx [7]. CP community could draw inspiration from success of using Transformers [59] for pre-training in NLP that was able to achieve significant milestones for a seemingly simple pre-training, and yet provide empirical performance guarantees with respect to model size and dataset size [30].

6. Concluding Remarks

In this work, we presented ARCH – the largest multiple instance captioning dataset for pathology images to date – carefully curated from textbooks and articles. We provided evidence to support our conjecture that pre-training on dense image captions in ARCH learns transferable representations for most CP tasks. We showed that ARCH is of similar complexity to its computer vision counterpart. We demonstrated the potential of new, yet unexplored sources of dense supervision in medical imaging, a field that is limited by the availability of high quality annotations. We posit that a model pre-trained on ARCH could be used without fine-tuning in multi-center screening, image retrieval, multiple instance or neural compression tasks and perform better than commonly used encoders trained on ImageNet. We are releasing ARCH[5] along with the datasets and models used in this work, such that the community could build on our work and fully explore the pre-training schemes that combine all of the available CP data.

[5] At https://warwick.ac.uk/TIA/data/ARCH (CC-NC).

References

[1] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[2] N. Alsubaie, M. Shaban, D. Snead, A. Khurram, and N. Rajpoot. A multi-resolution deep learning framework for lung adenocarcinoma growth pattern classification. In *Annual Conference on Medical Image Understanding and Analysis*, pages 3–11. Springer, 2018.

[3] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, and M. Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):1–11, 2018.

[4] R. Awan, K. Sirinukunwattana, D. Epstein, S. Jefferyes, U. Qidwai, Z. Aftab, I. Mujeeb, D. Snead, and N. Rajpoot. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports*, 7(1):1–12, 2017.

[5] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

[6] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[7] L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking*, 13(5):311–319, 2015.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[9] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

[10] R. Colling, H. Pitman, K. Oien, N. Rajpoot, P. Macklin, C.-P. A. in Histopathology Working Group, V. Bachtiar, R. Booth, A. Bryant, J. Bull, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *The Journal of pathology*, 249(2):143–150, 2019.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] K. Desai and J. Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] A. Ettinger, A. Elgohary, and P. Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[15] M. Fraz, S. Khurram, S. Graham, M. Shaban, M. Hassan, A. Loya, and N. Rajpoot. Fabnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. *Neural Computing and Applications*, pages 1–14, 2019.

[16] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

[17] J. Gamper, N. A. Koohbanani, K. Benet, A. Khuram, and N. Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pages 11–19. Springer, 2019.

[18] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[19] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019.

[20] P. Goyal, D. Mahajan, A. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6391–6400, 2019.

[21] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot. Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[23] N. Hegde, J. D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C. J. Cai, M. B. Amin, C. H. Mermel, et al. Similar image search for histopathology: Smily. *NPJ digital medicine*, 2(1):1–9, 2019.

[24] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15663–15674, 2019.

[25] M. S. Hosseini, L. Chan, G. Tse, M. Tang, J. Deng, S. Norouzi, C. Rowsell, K. N. Plataniotis, and S. Damaskinos. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11747–11756, 2019.

[26] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

[27] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[28] S. Javed, A. Mahmood, M. M. Fraz, N. A. Koohbanani, K. Benes, Y.-W. Tsang, K. Hewitt, D. Epstein, D. Snead, and N. Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis*, page 101696, 2020.

[29] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[30] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[31] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, 2020.

[32] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), 2019.

[33] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

[34] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

[35] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[36] M. Köbel, S. E. Kalloger, P. M. Baker, C. A. Ewanowich, J. Arseneau, V. Zherebitskiy, S. Abdulkarim, S. Leung, M. A. Duggan, D. Fontaine, et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. *The American journal of surgical pathology*, 34(7):984–993, 2010.

[37] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? arxiv e-prints, art. *arXiv preprint arXiv:1805.08974*, 2018.

[38] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[39] Q. Le, T. Sarlós, and A. Smola. Fastfood-computing hilbert space expansions in loglinear time. In *International Conference on Machine Learning*, pages 244–252, 2013.

[40] G. Lee, R. W. Veltri, G. Zhu, S. Ali, J. I. Epstein, and A. Madabhushi. Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *European urology focus*, 3(4-5):457–466, 2017.

[41] C. Li, H. Farkhoor, R. Liu, and J. Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[42] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[43] P. Liu, M. Lyu, I. King, and J. Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019.

[44] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, and F. Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019.

[45] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *arXiv preprint arXiv:2004.09666*, 2020.

[46] M. Y. Lu, M. Zhao, M. Shady, J. Lipkova, T. Y. Chen, D. F. Williamson, and F. Mahmood. Deep learning-based computational pathology predicts origins for cancers of unknown primary. *arXiv preprint arXiv:2006.13932*, 2020.

[47] R. Mormont, P. Geurts, and R. Marée. Multi-task pre-training of deep neural networks for digital pathology. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[48] K. Nagpal, D. Foote, Y. Liu, P.-H. C. Chen, E. Wulczyn, F. Tan, N. Olson, J. L. Smith, A. Mohtashamian, J. H. Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10, 2019.

[49] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 196–204. Springer, 2008.

[50] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[51] C. Resnick, Z. Zhan, and J. Bruna. Probing the state of the art: A critical look at visual representation evaluation. *arXiv preprint arXiv:1912.00215*, 2019.

[52] L. Roux, D. Racoceanu, F. Capron, J. Calvo, E. Attieh, G. Le Naour, and A. Gloaguen. Mitos & atypia. *Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep*, 1:1–8, 2014.

[53] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating

lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[54] M. Shaban, S. A. Khurram, M. M. Fraz, N. Alsubaie, I. Masood, S. Mushtaq, M. Hassan, A. Loya, and N. M. Rajpoot. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Scientific reports*, 9(1):1–13, 2019.

[55] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*, 2018.

[56] D. Tellez, D. Hoppener, C. Verhoef, D. Grunhagen, P. Nierop, M. Drozdzal, J. van der Laak, and F. Ciompi. Extending unsupervised neural image compression with supervised multitask learning. *arXiv preprint arXiv:2004.07041*, 2020.

[57] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[58] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[60] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[61] M. Veta, Y. J. Heng, N. Stathonikos, B. E. Bejnordi, F. Beca, T. Wollmann, K. Rohr, M. A. Shah, D. Wang, M. Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54:111–121, 2019.

[62] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[63] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.