# Learning Binary and Sparse Permutation-Invariant Representations for Fast and Memory Efficient Whole Slide Image Search

Sobhan Hemati†, Shivam Kalra†, Morteza Babaie†, H.R. Tizhoosh†, ‡

† Kimia Lab, University of Waterloo, Waterloo, ON, Canada
‡ Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
tizhoosh.hamid@mayo.edu

## Abstract

*Learning suitable Whole slide images (WSIs) representations for efficient retrieval systems is a non-trivial task. The WSI embeddings obtained from current methods are in Euclidean space not ideal for efficient WSI retrieval. Furthermore, most of the current methods require high GPU memory due to the simultaneous processing of multiple sets of patches. To address these challenges, we propose a novel framework for learning binary and sparse WSI representations utilizing a deep generative modelling and the* Fisher Vector. *We introduce new loss functions for learning sparse and binary permutation-invariant WSI representations that employ instance-based training achieving better memory efficiency. The learned WSI representations are validated on The Cancer Genomic Atlas (TCGA) and Liver-Kidney-Stomach (LKS) datasets. The proposed method outperforms* Yottixel *(a recent search engine for histopathology images) both in terms of retrieval accuracy and speed. Further, we achieve competitive performance against SOTA on the public benchmark LKS dataset for WSI classification.*

## 1. Introduction

The widespread adoption of digital pathology has spurred the digitization of tissue biopsy samples, known as whole slide images (WSIs) [2]. The computational pathology is expected to reduce the physicians' workload, improve diagnostic performance, and facilitate the teaching and research in pathology [32]. Deep learning is a successful tool for image analysis, including various applications in the medical domain. However, deep networks are challenging to adapt for WSI analysis [6]. These challenges include, but not limited to— tissue textures, rotationally invariant nature of the tissue, staining variations, and lack of fine-grained (patch-level) labelled data. Among all challenges, the major challenge is the sheer size of WSIs, typically $\gg 50{,}000 \times 50{,}000$ pixels. Furthermore, WSIs are arranged in a multi-resolution pyramidal structure containing images at different magnifications [32]. Therefore, memory and computationally efficient frameworks for WSI analyzing are an urgent need [23].

Patch extraction is typically the first step for the representation learning of a WSI. Commonly, thousands of representative patches can be extracted from a WSI. Processing the patches separately instead of the entire WSI eases the memory bottleneck; however, this leads to multi-vector embedding, which is non-trivial to transform to a single vector representation introducing new challenges, e.g., high data usage and compromised retrieval speed [19]. Computing a single-vector representation of a WSI is an active area of research [10, 18, 31]. Ideally, we are interested in a deep-learning solution that can be efficiently trained on WSI patches (at various magnifications), yielding a compact single-vector representation for the WSI, much more suitable for efficient retrieval tasks.

Multiple instance learning (MIL) enables learning on set data instead of using single instances during training. MIL is an appropriate method applicable to WSI representation, and as a result, there is a large body of papers exploring various MIL schemes for WSI representation learning [5, 10, 12, 14, 18, 27]. Although MIL has become a preferred method for WSI representation, it does have several limitations. First, the obtained euclidean embeddings cannot be directly used for WSI search in its raw form. Searching within large archives of WSIs through the nearest neighbour search would lead to a prohibitively large increase in memory demand and retrieval times [33]. As a result, the ancillary processing method is usually necessary to encode these embeddings into more suitable forms, i.e., binary and sparse embeddings facilitating the speed and memory efficiency in nearest neighbour search. Second, current MIL-based methods require all instances to be processed at once as a set (called *bag*), making it difficult to develop end-to-end training in a memory-efficient manner. Finally, current WSI engines, i.e., *Yottixel* [19], *SMILY* [9] ignore the a-

priori knowledge, such as tumor type, about WSIs for performing the search. It is ideal to employ all known attributes of WSIs for producing a more effective embedding. For this research, our contributions are as follows: 1) The compact (sparse and binary) and permutation-invariant WSI representations ideal for efficient WSI search in large archives, 2) the permutation-invariant representation of WSIs, trained end-to-end by feeding individual instances instead of a bag of instances which eases up the time and memory bottlenecks, enabling our methods to even incorporate patches at multiple magnification levels, 3) Learning representations guided by a-priori information, i.e., the tumor type as a way of self-supervision.

The rest of this paper is organized as follows: In Sec. 2 we briefly review the current related works on WSI representation learning. Then, in Sec. 3 we provide the details of our proposed framework. Next, in Sec. 4 we validate the effectiveness of our approach for search and classification tasks on two publicly available benchmark datasets. Finally, we conclude the paper in the Sec. 5.

## 2. Related works

In this section, we review the related literature on WSI representation learning. We organized the related literature into three main themes, i.e., heuristic deep architectures, Muti-instance Learning (MIL)-based methods, and dictionary learning approaches.

**Heuristic architectures.** These methods generally split the task into multiple separate steps to simplify the problem. First, there is an instance-based training where instances are smaller parts of WSIs, typically patches. Then, another network is trained to obtain WSI embeddings while capturing the spatial relationship between patches. For example, Bejnordi et al. [1] proposed to employ two sub-networks for processing high and low-resolution information separately and then attaching two networks together. Other works in this category are Spatio-Net [21] and the neural compression scheme proposed by Tellez et al. [31]. In Spatio-Net, a grid of embeddings for each patch and its neighbours are obtained by a CNN feature extractor, and then they are processed by 2D-LSTM layers to capture the spatial information. Tellez et al. [31] proposed a two-stage neural compression where the first stage is devoted to unsupervised representation learning of grid of all image patches per WSI. Then, they employed this trained model to obtain compressed patches and WSI. Finally, in one recent work, authors in [22] proposed a framework to choose between low and high-resolution information for WSI classification.

**Multiple instance learning (MIL).** Representing each WSI as a bag of patches makes MIL-based schemes a natural approach for end-to-end WSI representation learning [5, 14, 15, 18, 27]. One of the early works in MIL-based WSI classification was conducted by Hou et al. [12] where they first trained a patch level classifier and then a fusion model using MIL scheme to achieve WSI classification. In fact, one can regard this approach as a two-step instance-based MIL method where an algorithm determines instance classes. Motivated by this, Chikontwe et al. [3] proposed an end-to-end MIL-based method for simultaneous patch and WSI representation learning in a single framework where a center loss is introduced to map patch embeddings from the same WSI to a single centroid. Their approach achieved promising results compared with other MIL-based methods, especially two-stage MIL methods. Other recent MIL methods include [14] and [10] where pooling layers based on attention mechanism and Deep Sets [35] have been proposed. Finally, Kalra et al. [17] employed focal factor learning to modulate the aggregated patch-level predictions.

**Dictionary learning.** Another approach that can be used for the WSI representation is the *bag of visual words* (BoVW) [4] for encoding local image descriptors into one embedding. A more advanced version of BoVW that captures higher-order statistics to obtain the set representation is based on the *Fisher Kernel* theory and generative models [16]. Authors in [24] introduced Gaussian mixture model (GMM)-based Fisher Vector which can be calculated using the normalized gradient of the log-likelihood of the GMMs with respect to parameters, such as mixing coefficients, means, and variances, given a set of observations. Further, recently there has been some research to extract Fisher Vector from deep generative models [26, 36]. Although this set encoding ability makes Fisher Kernel a natural candidate for WSI representation learning. There are only a few papers that use Fisher theory and dictionary learning in general for the WSI representation task [29, 30, 37]. The reason could be attributed to the fact that Fisher Vector is formulated in a fully unsupervised manner using GMMs. However, considering the challenges inherent to pathology images (e.g., complex textures and colour variations), employing available WSI information, i.e., tumor type and primary diagnosis, in obtaining an efficient global representation is necessary. Further, GMM-based Fisher Vector captures no more than second-order statistics of data for set encoding. Besides, the training of GMMs is sub-optimal and not end-to-end. Finally, the obtained encodings are generally high-dimensional embeddings in Euclidean space, which are less desirable for WSI search due to their increased computation times for the distance computation.

## 3. Method

This section presents the proposed framework for learning compact WSI representations. First, we briefly review the relevant concepts, i.e., Fisher Kernel [16] and Fisher Vector theories [24]. Next, we describe the proposed method based upon variational autoencoders (VAEs)

and Fisher Vector theory. The proposed method is memory efficient during training and learns representations that are permutation-invariant, compact (sparse/binary), and can be conditioned on known information (e.g., the given tumor type) for the self-supervision. The proposed method is trained in an end-to-end manner on individual instances instead of a bag of instances to obtain representations for both patches and the WSI in its entirety.

### 3.1. Preparation

The key idea behind Fisher Kernel is to derive the kernel function from a generative probability model. Initially, the main motivation for deriving such kernels was bridging the gap between generative and discriminative models [16]: "the gradient of the log-likelihood with respect to a parameter describes how that parameter contributes to the process of generating a particular example". As a result, to take advantage of generative models in discriminative tasks, Jaakkola and Haussler proposed to employ the gradient space of the generative models to use the generative process as a similarity metric between examples (or set of examples, i.e., $\mathbf{X} = \{\mathbf{x}_t, t = 1, \dots, T\}$ where $T$ is the number of examples in the set) [16]. Let us consider a class of probability models $p(\mathbf{X} \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a parameter vector and $\mathbf{X}$ is set of examples, i.e., $\mathbf{X} = \{\mathbf{x}_t, t = 1, \dots, T\}$. The Fisher Score is then defined as

$$U_{\mathbf{X}} = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{X} \mid \boldsymbol{\theta}), \tag{1}$$

where the $U_{\mathbf{X}} \in \mathbb{R}^d$. The dimensionality $d$ of the Fisher Score is equal to the number of parameters in the generative model $p(\mathbf{X} \mid \boldsymbol{\theta})$ independent of the number of data points in the set $T$. The Fisher information matrix (FIM) is

$$\mathbf{I} = E_{\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})}\{U_{\mathbf{X}} U_{\mathbf{X}}^T\}. \tag{2}$$

Subsequently, the Fisher Kernel can be defined as

$$K(\mathbf{X}, \mathbf{Y}) = U_{\mathbf{X}}^T \mathbf{I}^{-1} U_{\mathbf{Y}} \tag{3}$$

Fisher Kernel can be used to calculate the similarity between two sets of data points [16]. Authors in [24] proposed the GMM-based Fisher Vector as a way to encode a set of local descriptors in a single embedding where the Fisher Vector is the normalized Fisher Score ($\mathbf{s}_F$) calculated as

$$\mathbf{s}_F = \frac{1}{T} \mathbf{L} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{X} \mid \boldsymbol{\theta}) = \mathbf{L} \frac{1}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_t \mid \boldsymbol{\theta}), \tag{4}$$

where $\mathbf{L}$ is calculated from Cholesky decomposition of inverse FIM, i.e., $\mathbf{I}^{-1} = \mathbf{L}^T \mathbf{L}$, with that assumption that data points in $\mathbf{X}$ are statistically independent.

### 3.2. Deep Compact Fisher Vector

The GMM-based Feature Vector exhibits shortcomings such as lack of employing available information, ignoring higher-order statistics in set encoding, its sub-optimal optimization, non-end-to-end training scheme, and the fact that Euclidean embeddings are not feasible for large archives. Motivated to remove these limitations, we propose a new type of Fisher Vector based on the deep generative models for the WSI representation learning. The contributions of our method are as follows.

1. To capture higher-order statistics while learning the set representation, we propose to employ generative models VAE here [20] for WSI representation learning.

2. We add a classification loss to the training such that the available WSI level primary diagnosis labels are employed during the training of the VAE.

3. We design the VAE to be conditioned on available information, e.g., the tumor type. Given the fact that every tumor type has its own specific cancer subtypes, this conditioning is expected to improve the quality of WSI embeddings.

4. More importantly, we propose two novel loss functions for compact (sparse and binary) and permutation-invariant WSI representation learning.

We start by training a VAE and then modify the VAE to be conditioned on tumor type. We add a classification loss to the end of the encoder part such that primary diagnosis label information is injected into the model space. Finally, we propose two novel loss functions for learning sparse and binary permutation-invariant representations.

**VAE loss function –** To learn the encoder and decoder parameters of the VAE, i.e., $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, that models distribution of $\mathbf{x}$, we assume the prior distribution on the random variable $\mathbf{z}$ is $p_{\boldsymbol{\theta}}(\mathbf{z})$ and as a result, $\mathbf{x}_t$ is sampled from $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$. In this case, one can show that the lower bound for the $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ can be calculated as

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq -q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{z}) + E_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right], \tag{5}$$

where $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ is the approximate posterior with parameters $\boldsymbol{\phi}$. The lower bound in Eq. 5 is known as variational lower bound and on the patch $\mathbf{x}_t$ and it is represented with $\mathcal{LB}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}_t)$. We aim to maximize $\mathcal{LB}$ to learn the generative model parameters. In context of the VAE model, $q_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{x}_t)$ and $p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{z}_t)$ are encoder and decoder, respectively. In order to learn the encoder and decoder parameters, i.e., $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, first we assume the prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}_t)$ is $\mathcal{N}(\mathbf{z}_t; 0, \mathbf{I})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{x}_t)$ and $p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{z}_t)$ follow the normal distributions $\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\mathbf{z}_t}, \boldsymbol{\sigma}^2_{\mathbf{z}_t} \mathbf{I})$ and $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\mathbf{x}_t}, \boldsymbol{\sigma}^2_{\mathbf{x}_t} \mathbf{I})$.

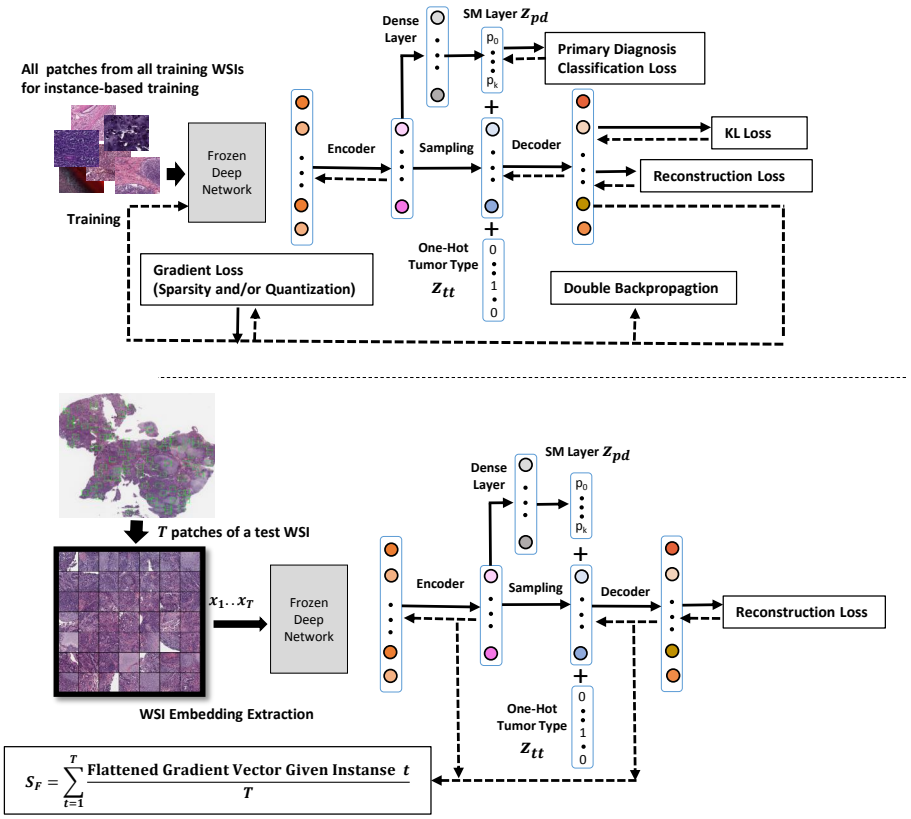


Figure 1. The first row represents the proposed architecture and associated instance-based training scheme. The second row shows the procedure for obtaining the WSI embedding for a set of patches of this WSI given the trained model.

In this case, by estimating the latent code using one-step Monte Carlo, the variational lower bound is

$$\mathcal{LB}(\boldsymbol{\phi},\boldsymbol{\theta},\mathbf{x}_t) = \log p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{z}_t) + \frac{1}{2}\sum_{j=1}^{d}(1+\log \boldsymbol{\sigma}^2_{\mathbf{z}_t(j)}) - \frac{1}{2}\|\boldsymbol{\mu}_{\mathbf{z}_t}\|^2 - \frac{1}{2}\|\boldsymbol{\sigma}_{\mathbf{z}_t}\|^2, \tag{6}$$

where $\mathbf{z}_t$ is sampled from $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_t}, \boldsymbol{\sigma}^2_{\mathbf{z}_t}\mathbf{I})$.

**Conditioned VAE –** As the tumor type of a WSI is always available, we conditioned VAE on the tumor type of the given WSI to draw benefit from this apriori knowledge. Following the [28], let's represent the tumor type as a one-hot encoded vector $\mathbf{z}_{tt}$, then we concatenate this vector to the $\mathbf{z}_t$. Furthermore, to inject WSI-level primary diagnosis information into the generative model, we add a classification loss to the last layer of the encoder before the sampling layer. We assign the WSI label to all patches extracted from that WSI. Then, we concatenate the softmax of predicted primary diagnosis $\mathbf{z}_{pd}$, with length $k$, to the latent space. The latent space associated with $\mathbf{x}_t$ that is fed to decoder is modified as $\mathbf{z}_t \leftarrow [\mathbf{z}_t, \mathbf{z}_{tt}, \mathbf{z}_{pd}]$. Considering the classification loss, so far, the loss function for training the conditioned VAE (CVAE) has the form

$$\mathcal{L}_{\text{CVAE}} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{kl} + \lambda_3\mathcal{L}_{cls}, \tag{7}$$

where minimizing the first two terms is equivalent to maximizing the variational lower bound, and the third loss is the classification loss of predicting cancer subtypes.

**Deep Sparse Fisher Vector –** Now, we propose a novel method for learning Sparse Fisher Vector (SFV). As the gradient space represents the WSI, we encourage sparsity in the gradient by adding the $l_1$ norm of the gradient of the loss function in Eq.7 to the overall training loss. To regularize the gradient, we utilize the *double backpropagation*, where given a batch of data points $\mathbf{X}$; the loss function can be written as

$$\mathcal{L}_{\text{SFV}} = \mathcal{L}_{\text{CVAE}} + \lambda_4 \sum_{\mathbf{W}_i \in \mathbb{W}} \|\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W}, \mathbf{X})\|_1, \tag{8}$$

where $\mathbb{W}$ is the set of CVAE parameters for all layers, $\mathbf{W}_i$ and $\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W}, \mathbf{X})$ are the is parameters and the gradient of the CVAE loss with respect to the $i^{th}$ layer parameters. To the best of our knowledge, such an end-to-end

Sparse Fisher Vector learning does not exist in the literature.

**Deep Binary Fisher Vector –** For learning deep binary Fisher Vector (BFV), inspired by the quantization-based learning in hashing literature [8, 11], we propose to reduce the quantization loss of the gradient of the CVAE loss with respect to each layer's parameters. We propose to find

$$\begin{aligned} \underset{\mathbf{B}_i, \nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X})}{\arg\min} \quad & \sum_{\mathbf{W}_i \in \mathbb{W}} \|\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X}) - \mathbf{B}_i\|_2^2 \\ \text{s.t.} \quad & \mathbf{B}_i \in \{-1,1\}^{d_i \times 1}, \end{aligned} \tag{9}$$

where $\mathbf{B}_i$ is the flattened binary representation of the gradient of the CVAE loss w.r.t parameters of the $i^{th}$ layer. In this case, the loss function to obtain BFV can be written as

$$\mathcal{L}_{\text{BFV}} = \mathcal{L}_{\text{CVAE}} + \lambda_5 \sum_{\mathbf{W_i} \in \mathbb{W}, \mathbf{B_i} \in \mathbb{B}} \|\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X}) - \mathbf{B}_i\|_2^2, \tag{10}$$

where $\mathbb{B}$ is the set of closest hamming vertices to gradients w.r.t all layers. Given the binary optimization variable $\mathbf{B}_i$, on each epoch, we employ the coordinate descent approach and update each of $\mathbf{B}_i$ and $\mathbf{W}_i$ while the other is fixed. For the case that $\mathbf{W}_i$ is fixed the problem turns to

$$\begin{aligned} \underset{\mathbf{B}_i}{\arg\min} \quad & \|\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X}) - \mathbf{B}_i\|_2^2 \\ \text{s.t.} \quad & \mathbf{B}_i \in \{-1,1\}^{d_i \times 1}, \end{aligned} \tag{11}$$

where by expanding Eq.11 it turns out the above minimization is equivalent to

$$\begin{aligned} \underset{\mathbf{B}_i}{\arg\max} \quad & \mathbf{B}_i^T . \nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{B}_i \in \{-1,1\}^{d_i \times 1}. \end{aligned} \tag{12}$$

This problem has the following closed-form solution [8]:

$$\mathbf{B}_i = sgn(\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X})). \tag{13}$$

The loss function can be for fixed $\mathbf{B}_i$ as

$$\mathcal{L}_{\text{BFV}} = \mathcal{L}_{\text{CVAE}} + \lambda_5 \sum_{\mathbf{W_i} \in \mathbb{W}} \|\nabla_{\mathbf{W}_i}\mathcal{L}_{\text{CVAE}}(\mathbb{W},\mathbf{X}) - \mathbf{B}_i\|_2^2 \tag{14}$$

This is similar to SFV learning in Eq. 8. The variables can be updated using double backpropagation.

**Deep Sparse Binary Fisher Vector –** Knowing that the length of obtained WSI embeddings is equal to the number of parameters in the generative model, we may be interested in compact (short) binary codes for more efficient WSI retrieval. We propose to employ both gradient sparsity and gradient quantization losses to achieve *Conditioned Sparse Binary Fisher Vector* (C-Deep-SBFV). Gradient sparsity pushes the generative model to use fewer parameters to generate a data point. As a result, the quality of embedding will be more robust to dropping some dimensions, i.e., gradient w.r.t some parameters of VAE. To choose effective dimensions for each tumor type we find the top $M$ parameters that provide the highest variance in their respective gradient values for the training data.

**VAE Architecture and Training Scheme –** The architecture of the proposed conditioned VAE is given in the first half of Fig. 1. We employed a frozen pre-trained CNN (DenseNet-121 [13]) as the backbone of the VAE. Each encoder and decoder parts contain three fully connected layers. The last layer of the encoder is fed to a softmax layer (SM Layer in Fig. 1) for primary diagnosis prediction. In order to condition the VAE, for each patch, the output of the softmax layer along with a one-hot encoded vector representing the available tumor type information of the patch is concatenated to the latent vector to create the $\mathbf{Z}_t$. Then, this vector is fed to the decoder part. As it can be seen from the Fig. 1, the CVAE is trained on a per-instance basis enabling to include even patches from multiple magnifications.

**WSI Embedding Extraction –** After the training phase, to obtain a single embedding for a WSI, all patches of that WSI are fed to the CVAE (see the second half of Fig. 1). Then, given the reconstruction loss, we calculate the average gradient over all patches using backpropagation to obtain the Fisher Score ($\mathbf{s}_F$). Based on Fisher Theory, we also need $\mathbf{L}$ obtained from FIM to normalize the vector and derive the Fisher Vector. However, given the computational load of calculating $\mathbf{L}$, we replace this with identity matrix and normalize the gradient using power and $l_2$ normalization steps proposed by [25]. In other words, representing the power and $l_2$ normalization steps as $\mathcal{S}(\cdot)$ operator, the conditioned deep compact Fisher Vector $\mathbf{v}_F$ is calculated from the Fisher Score $\mathbf{s}_F$:

$$\mathbf{v}_F = \mathcal{S}\left(\frac{1}{T}\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \|\mathbf{x}_t - \hat{\mathbf{x}}_t(\boldsymbol{\theta},\boldsymbol{\phi})\|_2^2\right) = \mathcal{S}\left(\mathbf{s}_F\right). \tag{15}$$

where $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t(\boldsymbol{\theta},\boldsymbol{\phi})$ are the patch embedding and its reconstruction. The size of the proposed feature vector is equal to the number of parameters in CVAE. The test-time, the one-hot vector of the tumor type, will be fed to the CVAE as a known parameter while the $\mathbf{z}_{pd}$ is calculated by the classifier.

## 4. Results

We evaluate the quality of the WSI embeddings obtained by the proposed method for both search and classification tasks. The datasets we employed are diagnostic slides from The Cancer Genomic Atlas (TCGA) repository [34] and the Liver-Kidney-Stomach (LKS) immunofluorescence [22] to conduct experiments.

| Site | Subtype | $n_{\text{slides}}$ | Yottixel | C-GMM-FV | C-Deep-FV | C-Deep-SFV | C-Deep-BFV | C-Deep-SBFV |
|---|---|---|---|---|---|---|---|---|
| Brain | LGG | 323 | 86.60 | 85.35 | 92.83 | 93.07 | 93.10 | 93.43 |
| | GBM | 387 | 88.68 | 87.63 | 93.44 | 93.76 | 93.99 | 94.24 |
| Endocrine | THCA | 198 | 97.98 | 97.02 | 98.74 | 99.24 | 99.24 | 98.50 |
| | ACC | 93 | 93.68 | 91.01 | 94.62 | 95.08 | 94.62 | 95.13 |
| | PCPG | 70 | 92.53 | 87.14 | 91.97 | 92.95 | 90.64 | 91.97 |
| Gastro. | ESCA | 55 | 60.95 | 58.82 | 72.00 | 65.42 | 75.43 | 66.66 |
| | COAD | 174 | 72.62 | 71.95 | 72.72 | 74.93 | 74.44 | 76.42 |
| | STAD | 157 | 79.75 | 79.75 | 78.59 | 80.75 | 83.48 | 83.97 |
| | READ | 61 | 24.24 | 29.62 | 23.52 | 31.77 | 30.30 | 31.37 |
| Gynaeco. | UCS | 37 | 65.62 | 66.66 | 78.26 | 72.13 | 74.62 | 73.01 |
| | UCEC | 206 | 84.23 | 82.82 | 89.31 | 87.29 | 83.33 | 88.16 |
| | CESC | 113 | 71.71 | 76.10 | 86.36 | 81.44 | 70.47 | 81.65 |
| | OV | 42 | 64.78 | 68.42 | 76.74 | 83.95 | 77.33 | 80.95 |
| Haematopoietic. | THYM | 80 | 93.41 | 93.49 | 93.56 | 93.97 | 96.93 | 94.04 |
| | DLBC | 14 | 47.61 | 42.10 | 35.29 | 54.54 | 80.00 | 50.00 |
| Liver, panc. | CHOL | 17 | 32.00 | 38.46 | 40.00 | 48.27 | 41.66 | 32.00 |
| | LIHC | 146 | 94.31 | 93.55 | 92.61 | 93.91 | 94.00 | 94.38 |
| | PAAD | 65 | 93.93 | 91.85 | 92.18 | 94.65 | 93.93 | 93.75 |
| Melanocytic malignancies | SKCM | 184 | 96.08 | 98.37 | 98.11 | 98.37 | 98.92 | 98.92 |
| | UVM | 40 | 76.92 | 92.30 | 90.90 | 92.30 | 94.73 | 94.73 |
| Prostate/testis | PRAD | 176 | 98.56 | 98.00 | 99.42 | 98.55 | 99.14 | 99.14 |
| | TGCT | 112 | 97.79 | 96.88 | 99.11 | 97.81 | 98.67 | 98.66 |
| Pulmonary | LUAD | 218 | 67.44 | 74.77 | 77.96 | 79.12 | 74.88 | 79.13 |
| | LUSC | 198 | 67.75 | 70.02 | 71.65 | 72.95 | 72.04 | 76.14 |
| | MESO | 27 | 7.14 | 43.24 | 50.00 | 51.28 | 40.00 | 31.25 |
| Urinary tract | BLCA | 193 | 90.41 | 88.26 | 92.34 | 92.83 | 94.20 | 95.93 |
| | KIRC | 195 | 88.88 | 88.26 | 91.82 | 93.75 | 91.47 | 93.81 |
| | KIRP | 142 | 77.73 | 75.53 | 82.97 | 84.01 | 84.05 | 84.78 |
| | KICH | 47 | 79.06 | 84.78 | 86.36 | 89.58 | 89.36 | 87.50 |

Table 1. F1-measure (in %) for majority-3 search through $k$-NN of the vertical search among 3770 test WSIs for Yottixel, C-GMM-FV, C-Deep-FV, C-Deep-SFV, C-Deep-BFV, and C-Deep-SBFV. Best F1-measure values highlighted.

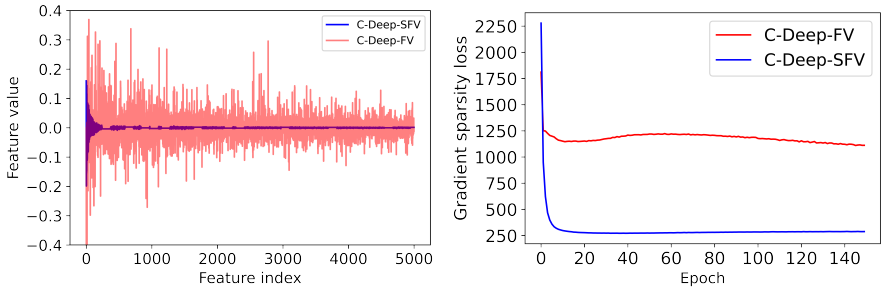


Figure 2. Left: Feature values across first 5,000 high variance dimensions for a WSI using C-Deep-SFV and C-Deep-FV. Right: Gradient sparsity loss ($l_1$ norm of the loss function gradient) of C-Deep-SFV and C-Deep-FV during the training epochs.

**WSI search –** For this experiment, we randomly selected 40% of the TCGA diagnostic WSIs as a test set and the rest for training. For both test and training WSIs, 15% of patches with $1000 \times 1000$ patch size have been selected based on the Yottixel (the same clustering method has been applied) [19]. Similar to [19], the vertical search has been applied on the test set (3,761 WSIs), and leave-one-out patient performed for searching WSIs through the same primary site. The majority of the top 3 similar cases have been used for predicting each query cancer subtype. Tab. 1 compares F1-measure between Yottixel, Conditional GMM-based Fisher Vector (C-GMM-FV), C-Deep-FV, C-Deep-SFV, C-Deep-BFV, and C-Deep-SBFV. Yottixel takes the median of minimum patch distances to calculate two WSIs dissimilarity, while C-GMM-FV and our proposed method obtain one embedding per WSI. Our proposed method improved the search F1-measure for all 29 cancer subtypes while the embeddings are binary and/or sparse. Although in almost all subtypes of two primary sites (Gynecological and Prostate/testis), C-Deep-FV performed better than other methods, almost in all cases, compact WSI embeddings obtained by gradient sparsity and quantization losses achieve even better search performance (see Tab. 1). The compactness of the proposed embeddings leads to high efficiency for WSI search in terms of memory usage and retrieval times. Fig. 2 (left) shows the embedding for C-Deep-FV and C-Deep-SFV across the first 5,000 high variance dimensions given the tissue type of the given WSI out of 1,407,105 parameters of our CVAE. Considering Fig. 2 (a), after encouraging sparsity on the gradients, the C-Deep-SFV can

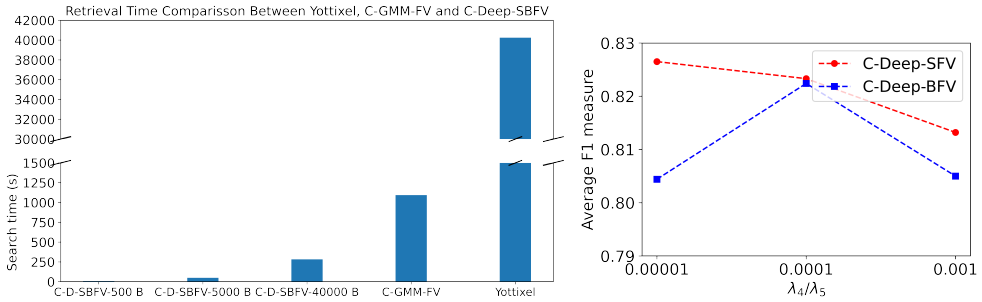


Figure 3. Left: Retrieval times for the leave-one-patient-out search (Tab. 1) for Yottixel and C-Deep-SBFV with different number of bits. Right: ablation study for gradient sparsity and quantization regularization terms. (full results in Tab. 4).

represent the WSI by significantly much fewer parameters leading to compact representations. Fig. 2 (right) shows the effectiveness of incorporating gradient sparsity loss in reducing the $l_1$ norm of the loss function gradient during the epochs.

The length of our proposed WSI embedding is equal to the number of trainable parameters of the generative model, which is 1,407,105. Although here we employed a relatively small generative model for models with millions of parameters, the embeddings may not be suitable for efficient WSI search. The proposed gradients sparsity solves this issue by enforcing the generative model to use a smaller number of parameters for generating samples. In other words, by imposing sparsity on gradients, one can argue that the parameters that the gradients are zero w.r.t them do not have a significant contribution to generating those samples, so they can be removed from embeddings. We have validated the effect of the sparsity loss by selecting the gradients w.r.t. a subset of some parameters that leads to high-variance gradients per tissue type. Tab. 2 shows the feature reduction results for the same search with keeping $500, 5000$, and $40,000$ high variance bits. Based on Tab. 2 and Fig. 3 (left), keeping even $4000$ high variance bits not only outperforms Yottixel and C-GMM-FV in terms of search performance but also leads to significantly faster search speed. Tab. 2 clearly shows that the sparsity term helps the network to produce embedding with fewer but more informative bits. The right column in most subtypes outperforms the left column. Based on the *Fisher Vector Theory* this is intuitive that the gradients w.r.t generative model parameters show the contribution of each parameter to generating a sample. More precisely, by encouraging sparsity on the gradients, fewer parameters are contributing to generation; consequently, more parameters can be dropped in the final embedding.

**WSI Classification –** For this task, we validated the quality of obtained WSI embeddings on LKS datasets. We trained a simple, fully connected network with two layers on top of the SFV embeddings for the purpose of WSI classification. The Liver-Kidney-Stomach (LKS) is the other publicly available dataset that we use for validating quality of WSI embeddings. The LKS dataset contains immunofluorescence WSIs realized by authors in [22]. The dataset contains 684 WSIs from four classes Anti-Mitochondrial Antibodies (AMA), Negative (Neg), Vessel-Type Anti-Smooth Muscle Antibodies (SMA-V), and Tubule-Type Anti-Smooth Muscle Antibodies (SMA-T). This dataset contains one low-resolution image and also a set of patches per WSI. Following the same split in [22], we compared C-Deep-SFV against the proposed method in this paper Selective Objective Switch (SOS), Reinforced Dynamic Multi-Scale (RDMS), [7] and three techniques for WSI classification, namely Image-Level, Patch-Level, and Conventional Multi-Scale (see [22] for more detail). For this experiment, we only employed low-resolution images for training the backbone and then for each WSI we used is low-resolution image along with 5% of high-resolution patches for training the CVAE and extracting the WSI embedding. The Tab. 3 presents the results in terms of precision, recall, F1, and accuracy measures where our method outperforms the Image-Level, Patch-Level, Conventional Multi-Scale, RDMS methods and achieves on par result compared with SOS. It is worth mentioning that our proposed architecture has one CNN backbone, while in SOS, two networks for low and high resolutions images have been used. Besides, embeddings obtained by or method are compact and suitable for WSI search.

**Ablation study –** To study how gradient sparsity and quantization loss may affect retrieval performance, we conducted comprehensive ablation experiments. Fig. 3 shows the average F1 measure across all sites for C-Deep-SFV and C-Deep-BFV where values $10^{-5}$, $10^{-4}$, and $10^{-3}$ have been tested for both $\lambda_4$ and $\lambda_5$. The average F1 decreases with increasing the $\lambda_4$. However, we should note that this experiment has been conducted with the same number of epochs (150). Our experiments showed that by increasing the $\lambda_4$ and also the number of epochs, the average F1 measure does not decrease. To see the effect of changing regu-

| Site | Subtype | 500 Bits | | 5000 Bits | | 40000 Bits | |
|---|---|---|---|---|---|---|---|
| | | C-Deep-BFV | C-Deep-SBFV | C-Deep-BFV | C-Deep-SBFV | C-Deep-BFV | C-Deep-SBFV |
| Brain | LGG | 84.48 | 89.02 | 86.27 | 90.93 | 93.21 | 94.06 |
| | GBM | 86.55 | 89.83 | 87.68 | 92.38 | 93.91 | 94.76 |
| Endocrine | THCA | 92.97 | 93.62 | 94.43 | 95.26 | 98.48 | 98.24 |
| | ACC | 82.87 | 84.37 | 86.91 | 93.19 | 94.68 | 96.25 |
| | PCPG | 73.43 | 75.40 | 69.49 | 84.61 | 91.30 | 92.64 |
| Gastro. | ESCA | 35.41 | 52.08 | 41.37 | 51.02 | 73.58 | 72.89 |
| | COAD | 64.37 | 70.96 | 68.16 | 70.96 | 77.65 | 74.58 |
| | STAD | 60.81 | 78.01 | 71.83 | 75.54 | 84.01 | 85.80 |
| | READ | 20.00 | 17.47 | 30.30 | 19.80 | 37.83 | 27.77 |
| Gynaeco. | UCS | 53.33 | 61.53 | 63.49 | 60.60 | 81.81 | 67.60 |
| | UCEC | 82.01 | 81.69 | 82.77 | 84.21 | 88.53 | 86.99 |
| | CESC | 64.03 | 64.48 | 63.73 | 73.87 | 82.24 | 82.35 |
| | OV | 64.93 | 52.17 | 74.66 | 70.42 | 82.50 | 83.95 |
| Haematopoietic. | THYM | 91.76 | 94.04 | 92.39 | 91.56 | 94.54 | 91.01 |
| | DLBC | 22.22 | 50.00 | 23.52 | 36.36 | 60.86 | 28.57 |
| Liver, panc. | CHOL | 25.00 | 9.52 | 20.00 | 23.07 | 34.78 | 38.46 |
| | LIHC | 81.36 | 84.24 | 85.09 | 85.34 | 92.30 | 93.95 |
| | PAAD | 58.18 | 66.12 | 68.42 | 74.79 | 98.29 | 90.90 |
| Melanocytic malignancies | SKCM | 96.02 | 94.31 | 97.57 | 95.58 | 97.57 | 94.91 |
| | UVM | 78.87 | 79.16 | 88.31 | 81.39 | 88.31 | 80.85 |
| Prostate/testis | PRAD | 91.37 | 96.04 | 94.05 | 95.18 | 98.29 | 98.85 |
| | TGCT | 86.84 | 93.69 | 90.58 | 92.37 | 97.32 | 98.24 |
| Pulmonary | LUAD | 62.30 | 67.89 | 69.27 | 68.62 | 76.64 | 75.71 |
| | LUSC | 61.65 | 62.31 | 61.08 | 64.51 | 72.72 | 72.53 |
| | MESO | 12.90 | 50.90 | 25.80 | 50.00 | 52.63 | 65.11 |
| Urinary tract | BLCA | 75.88 | 82.95 | 81.42 | 82.84 | 96.12 | 94.84 |
| | KIRC | 77.47 | 82.46 | 77.87 | 85.78 | 93.93 | 91.83 |
| | KIRP | 60.00 | 62.94 | 60.60 | 63.67 | 88.64 | 85.71 |
| | KICH | 42.25 | 52.27 | 50.00 | 54.34 | 89.79 | 87.23 |

Table 2. F1-measure (in %) for majority-3 search through $k$-NN of the vertical search among 3,761 test WSIs for feature reduction effect by comparing top 500, 1000, and 5000 high variance features.

| | SMA-T class | | | Negative Class | | | AMA Class | | | SMA-V Class | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | F1 | PR | RE | F1 | PR | RE | F1 | PR | RE | F1 | PR | RE | ACC |
| Image-Level | 16.67 | 100.00 | 9.09 | 88.00 | 81.15 | 96.12 | 89.89 | 90.90 | 88.89 | 66.67 | 73.68 | 60.87 | 81.95 |
| Patch-Level | 00.00 | 00.00 | 00.00 | 79.67 | 68.53 | 95.15 | 84.71 | 90.00 | 80.00 | 23.53 | 36.36 | 17.39 | 69.27 |
| Multi-Scale | 47.06 | 66.67 | 36.36 | 90.83 | 86.09 | 96.12 | 87.06 | 92.50 | 82.22 | 77.78 | 79.55 | 76.09 | 85.37 |
| RDMS [7] | 55.56 | 71.43 | 45.45 | 93.00 | 95.88 | 90.29 | 91.49 | 87.76 | **95.56** | 83.67 | 78.85 | 89.13 | 88.78 |
| SOS [22] | **70.00** | 71.43 | **63.64** | **94.06** | **95.96** | 92.23 | 93.48 | 91.49 | **95.56** | **85.42** | 82.00 | **89.13** | **90.73** |
| D-SFV (Ours) | 66.67 | **85.71** | 54.55 | 93.84 | 91.67 | **96.12** | **94.51** | **93.48** | **95.56** | 84.44 | **86.36** | 82.61 | **90.73** |

Table 3. Comparison of different WSI classification methods against Deep-SFV on Liver-Kidney-Stomach (LKS) dataset. F1, PR, and RE are in %.

larization parameters in more detail, we refer the readers to Tab. 4 in the supplementary material.

## 5. Conclusions

We proposed a new framework based on deep conditional generative modelling and *Fisher Vector Theory* for compact WSI representation. Unlike the common practice for WSI representation, i.e., MIL scheme, the training for the proposed method is instance-based, and as a result, GPU memory usage is the same as conventional training. Furthermore, we introduced new loss functions, gradient sparsity and gradient quantization for learning sparse and binary permutation-invariant representations, namely C-Deep-SFV and C-Deep-BFV, suitable for efficient WSI retrievals. We showed that gradient sparsity loss function pushes the generative model to use parameters for generating a sample, and as a result, one can reduce the dimensionality of the WSI embeddings and still achieve a good performance. The WSIs representations were validated on the largest public archive of WSIs, The TCGA WSIs and also the LKS dataset for both WSI search and classification tasks. The proposed method outperforms *Yottixel* a recent search engine for histopathology images and *GMM-based Fisher Vector*. Furthermore, we also achieved competitive results against state-of-the-art in WSI classifications on both lung and LKS public benchmark datasets.

## 6. Data availability

The diagnostic slides from The Cancer Genomic Atlas (TCGA) repository [34] dataset is available here. The Liver

Kidney Stomach (LKS) [22] is available here in this repository.

## References


[1] Babak Ehteshami Bejnordi, Guido Zuidhof, Maschenka Balkenhol, Meyke Hermsen, Peter Bult, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, and Jeroen van der Laak. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging*, 4(4):044504, 2017. 2

[2] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019. 1

[3] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 519–528, Cham, 2020. Springer International Publishing. 2

[4] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[5] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 1, 2

[6] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019. 1

[7] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 317–325. Springer, 2018. 7, 8

[8] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929, 2013. 5

[9] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. Similar image search for histopathology: Smily. *NPJ digital medicine*, 2(1):1–9, 2019. 1

[10] Sobhan Hemati, Shivam Kalra, Cameron Meaney, Morteza Babaie, Ali Ghodsi, and Hamid Tizhoosh. Cnn and deep sets for end-to-end whole slide image representation learning. In *Medical Imaging with Deep Learning*, 2021. 1, 2

[11] Sobhan Hemati, Mohammad Hadi Mehdizavareh, Shojaeddin Chenouri, and Hamid R Tizhoosh. A non-alternating graph hashing algorithm for large scale image search. *arXiv preprint arXiv:2012.13138*, 2020. 5

[12] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016. 1, 2

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2

[15] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020. 2

[16] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*, pages 487–493, 1999. 2, 3

[17] Shivam Kalra, Mohammed Adnan, Sobhan Hemati, Taher Dehkharghanian, Shahryar Rahnamayan, and Hamid R Tizhoosh. Pay attention with focus: A novel learning scheme for classification of whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 350–359. Springer, 2021. 2

[18] Shivam Kalra, Mohammed Adnan, Graham Taylor, and Hamid R Tizhoosh. Learning permutation invariant representations using memory networks. In *European Conference on Computer Vision*, pages 677–693. Springer, 2020. 1, 2

[19] Shivam Kalra, HR Tizhoosh, Charles Choi, Sultaan Shah, Phedias Diamandis, Clinton JV Campbell, and Liron Pantanowitz. Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757, 2020. 1, 6

[20] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019. 3

[21] Bin Kong, Xin Wang, Zhongyu Li, Qi Song, and Shaoting Zhang. Cancer metastasis detection via spatially structured deep network. In *International Conference on Information Processing in Medical Imaging*, pages 236–248. Springer, 2017. 2

[22] Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, and Brian C Lovell. Sos: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3862–3871, 2020. 2, 5, 7, 8, 9

[23] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019. 1

[24] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007*

*IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2, 3

[25] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010. 5

[26] Zhaofan Qiu, Ting Yao, and Tao Mei. Deep quantization: Encoding convolutional activations with deep generative model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6759–6768, 2017. 2

[27] Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234, 2017. 1, 2

[28] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015. 4

[29] Yang Song, Hang Chang, Heng Huang, and Weidong Cai. Supervised intra-embedding of fisher vectors for histopathology image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–106. Springer, 2017. 2

[30] Yang Song, Ju Jia Zou, Hang Chang, and Weidong Cai. Adapting fisher vectors for histopathology image classification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 600–603. IEEE, 2017. 2

[31] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[32] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9, 2018. 1

[33] J. Wang, T. Zhang, J. Song, N. Sebe, and H. Shen. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(04):769–790, apr 2018. 1

[34] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 5, 8

[35] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[36] Shuangfei Zhai, Walter Talbott, Carlos Guestrin, and Joshua Susskind. Adversarial fisher vectors for unsupervised representation learning. *Advances in Neural Information Processing Systems*, 32:11158–11168, 2019. 2

[37] Shujin Zhu, Yuehua Li, Shivam Kalra, and Hamid R Tizhoosh. Multiple disjoint dictionaries for representation of histopathology images. *Journal of Visual Communication and Image Representation*, 55:243–252, 2018. 2

# Supplementary Material

## A. Detailed ablation study

The Table 4 shows the detailed ablation study on the gradient sparsity and quantization losses regularization parameters. More precisely, The F1 measure across all sites for C-Deep-SFV and C-Deep-BFV with regularization parameters equal to $10^{-5}$, $10^{-4}$, and $10^{-3}$for $\lambda_4$ and $\lambda_5$ are reported in Table 4.

## B. Full names for cancer subtypes

The full description of the abbreviations for cancer subtypes used in this paper have been presented in Table 5.

| Site | Subtype | F1-measure (in %) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\lambda_4 = 10^{-5}$ C-Deep-SFV | $\lambda_5 = 10^{-5}$ C-Deep-BFV | $\lambda_4 = 10^{-4}$ C-Deep-SFV | $\lambda_5 = 10^{-4}$ C-Deep-BFV | $\lambda_4 = 10^{-3}$ C-Deep-SFV | $\lambda_5 = 10^{-3}$ C-Deep-BFV |
| Brain | LGG | 90.88 | 92.87 | 93.07 | 93.10 | 92.72 | 93.80 |
| | GBM | 91.67 | 93.91 | 93.76 | 93.99 | 93.53 | 94.70 |
| Endocrine | THCA | 98.48 | 97.29 | 99.24 | 99.24 | 98.74 | 98.75 |
| | ACC | 92.55 | 91.30 | 95.08 | 94.62 | 95.13 | 94.62 |
| | PCPG | 88.40 | 90.07 | 92.95 | 90.64 | 91.42 | 90.37 |
| Gastro. | ESCA | 69.42 | 67.76 | 65.42 | 75.43 | 77.04 | 70.17 |
| | COAD | 73.38 | 75.97 | 74.93 | 74.44 | 76.21 | 74.01 |
| | STAD | 84.07 | 79.23 | 80.75 | 83.48 | 84.45 | 84.01 |
| | READ | 25.49 | 33.33 | 31.77 | 30.30 | 30.18 | 29.90 |
| Gynaeco. | UCS | 81.15 | 74.28 | 72.13 | 74.62 | 73.23 | 71.64 |
| | UCEC | 89.97 | 86.18 | 87.29 | 83.33 | 89.42 | 85.58 |
| | CESC | 82.72 | 84.30 | 81.44 | 70.47 | 86.84 | 74.17 |
| | OV | 79.48 | 73.68 | 83.95 | 77.33 | 79.01 | 81.01 |
| Haematopoietic. | THYM | 94.54 | 93.49 | 93.97 | 96.93 | 93.56 | 95.23 |
| | DLBC | 60.86 | 42.10 | 54.54 | 80.00 | 35.29 | 60.00 |
| Liver, panc. | CHOL | 41.66 | 25.00 | 48.27 | 41.66 | 14.81 | 16.66 |
| | LIHC | 94.91 | 92.66 | 93.91 | 94.00 | 88.66 | 92.35 |
| | PAAD | 91.97 | 90.90 | 94.65 | 93.93 | 86.82 | 93.12 |
| Melanocytic malignancies | SKCM | 98.13 | 98.37 | 98.37 | 98.92 | 97.23 | 97.86 |
| | UVM | 90.41 | 92.30 | 92.30 | 94.73 | 88.37 | 89.18 |
| Prostate/testis | PRAD | 99.42 | 99.14 | 98.55 | 99.14 | 98.56 | 98.85 |
| | TGCT | 99.11 | 98.66 | 97.81 | 98.67 | 98.56 | 98.23 |
| Pulmonary | LUAD | 77.84 | 75.84 | 79.12 | 74.88 | 76.95 | 78.70 |
| | LUSC | 72.67 | 74.20 | 72.95 | 72.04 | 71.72 | 72.20 |
| | MESO | 63.63 | 50.00 | 51.28 | 39.99 | 72.72 | 50.00 |
| Urinary tract | BLCA | 94.20 | 94.62 | 92.83 | 94.20 | 96.14 | 94.02 |
| | KIRC | 92.73 | 90.90 | 93.75 | 91.47 | 93.50 | 89.35 |
| | KIRP | 86.69 | 83.94 | 84.01 | 84.05 | 87.63 | 78.41 |
| | KICH | 90.52 | 90.32 | 89.58 | 89.36 | 90.72 | 87.64 |

Table 4. Ablation study on $\lambda_4$ and $\lambda_5$ regularization parameters based on Majority-3 search through $k$-NN of the vertical search among 3761 test WSIs.

| Abbreviation | Primary Diagnosis |
|---|---|
| ACC | Adrenocortical Carcinoma |
| BLCA | Bladder Urothelial Carcinoma |
| CESC | Cervical Squamous Cell Carcinoma and Endocervical Adenoc. |
| CHOL | Cholangiocarcinoma |
| COAD | Colon Adenocarcinoma |
| ESCA | Esophageal Carcinoma |
| GBM | Glioblastoma Multiforme |
| KICH | Kidney Chromophobe |
| KIRC | Kidney Renal Clear Cell Carcinoma |
| KIRP | Kidney Renal Papillary Cell Carcinoma |
| LGG | Brain Lower Grade Glioma |
| LIHC | Liver Hepatocellular Carcinoma |
| LUAD | Lung Adenocarcinoma |
| LUSC | Lung Squamous Cell Carcinoma |
| MESO | Mesothelioma |
| OV | Ovarian Serous Cystadenocarcinoma |
| PAAD | Pancreatic Adenocarcinoma |
| PCPG | Pheochromocytoma and Paraganglioma |
| PRAD | Prostate Adenocarcinoma |
| READ | Rectum Adenocarcinoma |
| STAD | Stomach Adenocarcinoma |
| TGCT | Testicular Germ Cell Tumors |
| THCA | Thyroid Carcinoma |
| UCS | Uterine Carcinosarcoma |

Table 5. Full description for primary diagnosis abbreviations used in the paper.