Title

Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications

Permalink

https://escholarship.org/uc/item/6n6147g5

Journal

IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5)

ISSN

0162-8828

Authors

Chang, Hang
Han, Ju
Zhong, Cheng
et al.

Publication Date

2018-05-01

DOI

10.1109/tpami.2017.2656884

Peer reviewed

Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications

Hang Chang*, *Member, IEEE,* Ju Han, *Member, IEEE,* Cheng Zhong, *Member, IEEE,* Antoine M. Snijders, Jian-Hua Mao

Abstract—The capabilities of (I) learning transferable knowledge across domains; and (II) fine-tuning the pre-learned base knowledge towards tasks with considerably smaller data scale are extremely important. Many of the existing transfer learning techniques are supervised approaches, among which deep learning has the demonstrated power of learning domain transferrable knowledge with large scale network trained on massive amounts of labeled data. However, in many biomedical tasks, both the data and the corresponding label can be very limited, where the unsupervised transfer learning capability is urgently needed. In this paper, we proposed a novel multi-scale convolutional sparse coding (MSCSC) method, that (I) automatically learns filter banks at different scales in a joint fashion with enforced scale-specificity of learned patterns; and (II) provides an unsupervised solution for learning transferable base knowledge and fine-tuning it towards target tasks. Extensive experimental evaluation of MSCSC demonstrates the effectiveness of the proposed MSCSC in both regular and transfer learning tasks in various biomedical domains.

Index Terms—Transfer Learning, Sharable Information, Convolutional Sparse Coding, Deep Learning, Biomedical Application, Brain Tumors, Low Dose Ionizing Radiation (LDIR), Mouse Model, Breast Cancer Subtypes.

1 Introduction

Recent neuroscience findings [1], [2] have identified the complex hierarchy in the neocortex for the representation of observations. Motivated by these findings, one branch of the machine learning community has attempted to build information representations, through computational modules, which share similar properties with those in the neocortex. For the past decade, deep learning has gained momentum as a result of its demonstrated capability of improved performance for various automation tasks and its potential for future research. Among different deep learning approaches, Convolutional Neural Networks (CNNs) [3]–[7] and Deep Belief Networks (DBNs) [8], [9] are the most well-established techniques.

Along with the development of modern deep neural networks, one curious phenomenon exhibits that, regardless of natural image dataset or even training objectives [9]–[11], the first layer of the deep neural network always captures standard features that resemble either Gabor filters or color blobs. The common appearance of filters learned from the first few layers provides the domain adaptive/transferrable base knowledge, which can serve as the basis for transfer learning [12]–[14]. During transfer learning with deep neural networks, a base deep neural network is first trained on a base dataset and task, and the learned knowledge (e.g., features, representation) is then transferred to a target network to be trained on a target dataset and task. Typically, the first n layers of the target deep neural network is initialized as the first n layers of the base deep neural network; while with the remaining layers randomly initialized and trained towards the target dataset and task. Depending on the size of the target dataset and the size of the network (i.e., the number of parameters), the first n layers of the target deep neural network can either be frozen (i.e., remain unchanged during training on the new task), or be fine-tuned based on backpropagation strategy towards the new task, which is a balance between specificity and generality of derived knowledge. Although deep neural networks have been successfully applied in various biomedical tasks, the training of such large scale networks typically requires massive amounts of labeled data, which can be very limited in many biomedical tasks.

In this paper, we proposed a novel method, namely Multi-Scale Convolutional Sparse Coding (MSCSC), which automatically learns filter banks at different scales in a joint fashion with enforced scale-specificity, and therefore not only improves the classification performance on many biomedical tasks, but also provides an unsupervised solution for transfer learning.

The rest of this paper is organization as follows: Section 2 briefly reviews related studies. Section 3 describes the details of proposed MSCSC model. Section 4 and Section 5 elaborate the experimental design, followed by detailed discussion on the evaluation results. Lastly, section 6 concludes the paper.

- ** Correspondence should be addressed to Hang Chang (hchang@lbl.gov)*
- *Hang Chang and Ju Han and Cheng Zhong and Antoine M. Snijders and Jian-Hua Mao are with Berkeley Biomedical Data Science Center (BBDS: http://bbds.lbl.gov), Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A*
- *All related resources have been released for public consumption at BMIHub - http://bmihub.org*

This work was supported by NIH R01 CA184476 carried out at Lawrence Berkeley National Laboratory.

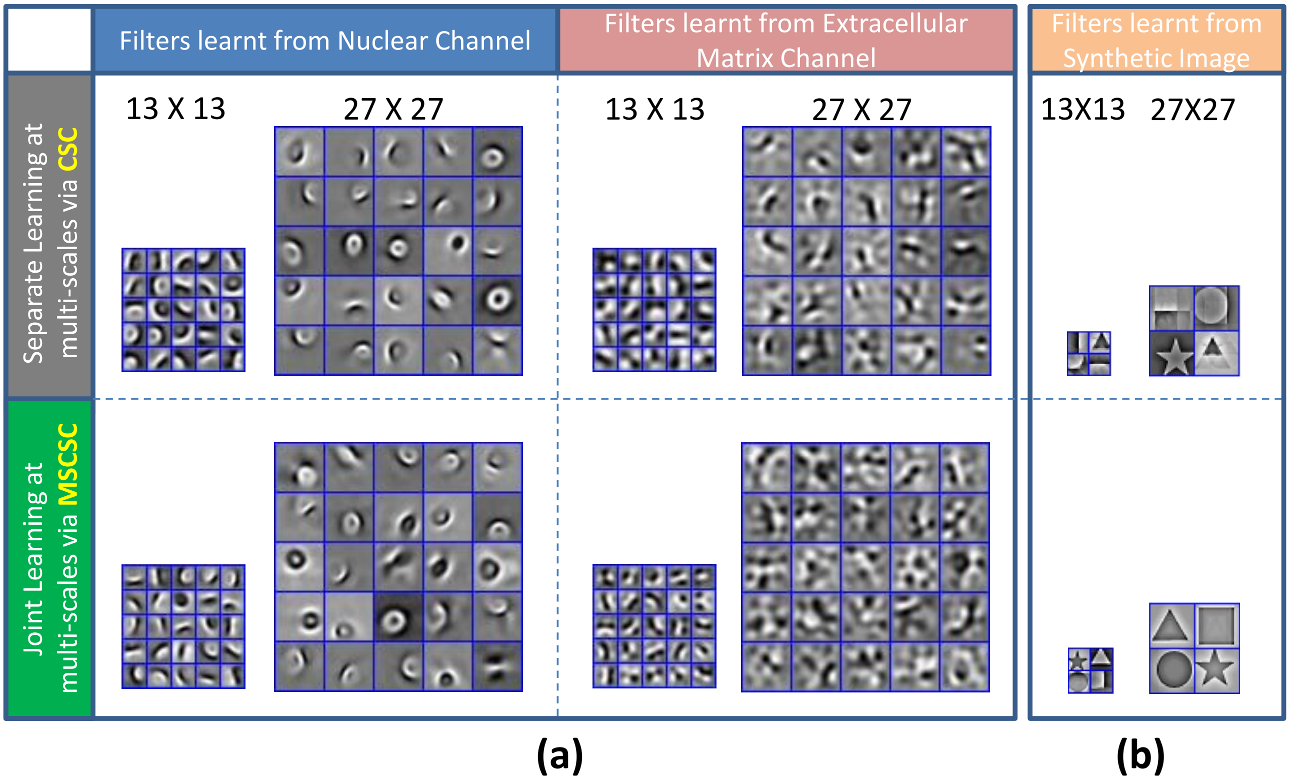

Filters learnt from Nuclear Channel
Filters learnt from Extracellular Matrix Channel
Filters learnt from Synthetic Image
Separate Learning at multi-scales via CSC
Joint Learning at multi-scales via MSCSC
13 X 13
27 X 27
13 X 13
27 X 27
13X13
27X27
(a)
(b)

Fig. 1. Comparison of multi-scale filters learnt via MSCSC and CSC from **(a)** GBM dataset, where each tissue image is decomposed into two spectra (channels) corresponding to nuclei and extracellular matrix (ECM) for filter learning; and **(b)** A synthetic image, consisting of four distinct binarized shapes (★, ■, ●, ▲) at two different scales (13×13 and 27×27). It is clear that, through joint learning via MSCSC, the filters at smaller scale (i.e., 13×13) mainly captures lower-level features/small objects (e.g., edges in GBM dataset and small shapes in synthetic image), while the filters at larger scale (i.e., 27×27) are more responsible for higher-level features/large objects (e.g., complex pattern in ECM of GBM dataset and large shapes in synthetic image). However, filters learnt separately per scale via CSC do not have such scale-specificity, and present a mixture of low-/high-level features at both scales, which might lead to feature redundancy across scales. It is also worth to mention that, for GBM dataset, the difference in scale-specificity becomes more distinct for filters learnt from ECM, since compared with nuclear chromatin, ECM sees much more complex patterns, which CSC fails to capture.

2 Related Work

In recent years, convolutional sparse coding has received increasing research interest in computer vision and machine learning communities [15]–[20], due mainly to its capability of learning shift-invariant filters with complex patterns.

Meanwhile, in the field of transfer learning via deep neural networks, recent studies [21]–[23] have shown that, given a target dataset which is significantly smaller than the base one, transfer learning can be a powerful tool to enable training a large target network to obtain state-of-the-art results in various tasks without over-fitting, which suggests that the first few layers of a deep neural network, trained on a large scale base dataset, can capture domain adaptive/transferable knowledge, which is fairly general at least in the natural image domain. Although deep neural networks have been successfully applied in various biomedical tasks, the training of such large scale networks typically requires massive amounts of labeled data, which can be very limited in many biomedical tasks.

3 Multi-Scale Convolutional Sparse Coding

In this section, we describe the proposed multi-scale convolutional sparse coding model. Without the loss of generality, we demonstrate MSCSC with 2D images as input. Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ be a training set containing N images with dimension $m \times n$. Let $\mathbf{D} = \{\mathbf{d}_{s,k}\}_{s=1,k=1}^{S,K}$ be the 2D multi-scale convolutional filter bank with S different scales, and K filters per scale, where each $\mathbf{d}_{s,k}$ is an $h_s \times h_s$ convolutional kernel. Define $\mathbf{Z} = \{\mathbf{Z}^i\}_{i=1}^{N}$ as the set of sparse feature maps, where $\mathbf{Z}^i = \{\mathbf{z}^i_{s,k}\}_{s=1,k=1}^{S,K}$ consists of $S \times K$ feature maps for the reconstruction of image $\mathbf{x}_i$. MSCSC aims to decompose each training image $\mathbf{x}_i$ as the sum of a series of sparse feature maps $\mathbf{z}^i_{s,k} \in \mathbf{Z}^i$ convolved with kernels $\mathbf{d}_{s,k}$ from the filter bank $\mathbf{D}$, by solving the following objective function:

$$\min_{\mathbf{D},\mathbf{Z}} \mathcal{L} = \sum_{i=1}^{N} \left\{ \left\| \mathbf{x}_i - \sum_{s=1}^{S} \sum_{k=1}^{K} \mathbf{d}_{s,k} * \mathbf{z}^i_{s,k} \right\|_F^2 + \alpha \sum_{s=1}^{S} \sum_{k=1}^{K} \left\| \mathbf{z}^i_{s,k} \right\|_1 \right\}$$
$$\text{s.t. } \|\mathbf{d}_{s,k}\|_2^2 = 1, \forall k = 1, \ldots, K; \forall s = 1, \ldots, S \qquad (1)$$

where the first and the second term represent the reconstruction error and the ℓ_1-norm penalty, respectively; α is a regularization parameter; $*$ is the 2D discrete convolution operator; and the filters are constrained to have unit energy to avoid trivial solutions. The construction of $\mathbf{D}$ is a balance between the reconstruction error and the ℓ_1-norm penalty.

Note that the objective of Eq. (1) is not jointly convex with respect to $\mathbf{D}$ and $\mathbf{Z}$, but is convex with respect to either one of them with the other fixed [24]. We thus solve Eq. (1) by optimizing $\mathbf{D}$ and $\mathbf{Z}$ in an alternative fashion, *i.e.,* iteratively performing the two steps that first compute $\mathbf{Z}$ and then updating $\mathbf{D}$. Specifically, we use the Iterative Shrinkage Thresholding Algorithm (ISTA) to solve for the sparse feature maps $\mathbf{Z}$; and use the stochastic gradient descent [15] for updating the convolutional dictionary $\mathbf{D}$. Alternative methods for updating the dictionary can be found in [16], [17], [20], and the proposed optimization procedure is sketched in Algorithm 1. It is clear that the proposed optimization procedure utilizes the standard ISTA strategy with indices over the different scales.

Algorithm 1 MSCSC Algorithm

```
Input: Training set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, $K$, $\alpha$
Output: Convolutional filter bank $\mathbf{D} = \{\mathbf{d}_{s,k}\}_{s=1,k=1}^{S,K}$
Initialize: $\mathbf{D} \sim \mathcal{N}(0,1)$, $\mathbf{Z} \leftarrow \mathbf{0}$
repeat
    for $i = 1$ to $N$ do
        Normalize each kernel in $\mathbf{D}$ to unit energy
        Fixing $\mathbf{D}$, compute sparse feature maps $\mathbf{Z}^i$ by solving
        $\mathbf{Z}^i \leftarrow \arg\min_{\mathbf{z}_{s,k}^i \in \mathbf{Z}^i} \|\mathbf{x}_i - \sum_{s=1}^{S}\sum_{k=1}^{K} \mathbf{d}_{s,k} * \mathbf{z}_{s,k}^i\|_{\mathrm{F}}^2 + \alpha \sum_{s=1}^{S}\sum_{k=1}^{K} \left\|\mathbf{z}_{s,k}^i\right\|_1$
        Fixing $\mathbf{Z}$, update $\mathbf{D}$ as
        $\mathbf{D} \leftarrow \mathbf{D} - \mu \nabla_{\mathbf{D}} \mathcal{L}(\mathbf{D}, \mathbf{Z})$
    end for
until Convergence (maximum iterations reached or objective function $\leq$ threshold)
```

4 Evaluation of MSCSC on Regular Classification Tasks

This section provides experimental evaluation of MSCSC on tissue histology classification and the classification of breast cancer subtypes, followed by detailed discussion on the experimental results.

4.1 Evaluation of MSCSC on Tissue Histology Classification

In this section, we present detailed experimental design and evaluation of MSCSC on the task of tissue histology classification. The corresponding classification pipeline, namely Multi-Scale-CSCSPM, was built upon MSCSC and SPM, and applied on two distinct tumor datasets, curated from The Cancer Genome Atlas (TCGA), namely (i) Glioblastoma Multiforme (GBM) and (ii) Kidney Renal Clear Cell Carcinoma (KIRC), which are publicly available from the NIH (National Institute of Health) repository.

4.1.1 Multi-Scale Multi-Spectral Feature Extraction for Tissue Histology Classification

As suggested in [25], different spectra of biomedical images usually capture distinct targets of interests, and applying CSC to each spectrum separately enables learning of biological-component-specific feature detectors, which helps improve the classification performance. Therefore, we adopt the same configuration as in [25], and apply the proposed MSCSC to two separate spectra produced through color decomposition [26], which characterize the nuclear chromatin and the extracellular matrix, respectively.

Without the loss of generality, we assume that the number of filters for each spectrum (channel) is K per scale, the number of scales is S, and the number of spectra (channels) is W after decomposition; the 2D feature map $\mathbf{y}_{s,k}^w$ is then defined as: $\mathbf{y}_{s,k}^w = \mathbf{d}_{s,k}^w * \hat{\mathbf{x}}^w$, for $1 \leq s \leq S$, $1 \leq k \leq K$ and $1 \leq w \leq W$, where $\hat{\mathbf{x}}^w$ is the w-th spectrum component of input image $\mathbf{x}$ and $\mathbf{d}_{s,k}^w \in \mathbf{D}^w$ is the k-th convolutional kernel at scale s in filter bank $\mathbf{D}^w$ learned over spectrum with index w.

The architecture for multi-scale multi-spectral tissue histology feature extraction is illustrated in Figure 2, which consists steps as follows,

1) Color decomposition (CoD). An input image is first decomposed and divided into two separate spectrum [26], corresponding to the nuclear chromatin and the extracellular matrix, respectively.
2) Multi-scale convolution. Each decomposed spectra is convolved with spectrum-specific multi-scale filters learnt via MSCSC.
3) Element-wise absolute value rectification (Abs). The Abs layer computes absolute value element-wisely in each feature map, $\mathbf{y}_{s,k}^w$, to avoid the cancelation effect in sequential operations.
4) Local contrast normalization (LCN). The LCN layer aims to enhance the stronger feature responses and suppress weaker ones across feature maps, $\{\mathbf{y}_{s,k}^w\}_{s=1,k=1}^{S,K}$, in each spectrum ($w$), by performing local subtractive and divisive operations [18], [27].
5) Max-pooling (MP). The MP layer partitions each feature map into non-overlapping windows and extracts the maximum response from each of the pooling window. It allows local invariance to translation [27].
6) Concatenation of features from each spectrum to form the multi-scale multi-spectral tissue features.

After extraction, the multi-scale multi-spectral tissue features, with dimensionality SKW, are fed into SPM framework for summarization and classification as described in the following section.

4.1.2 Feature Summarization via SPM

We adopt SPM to construct tissue morphometric context [25], [28]–[31] as the final representation for tissue classification. Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_T] \in \mathbb{R}^{SKW \times T}$ be the feature set of T

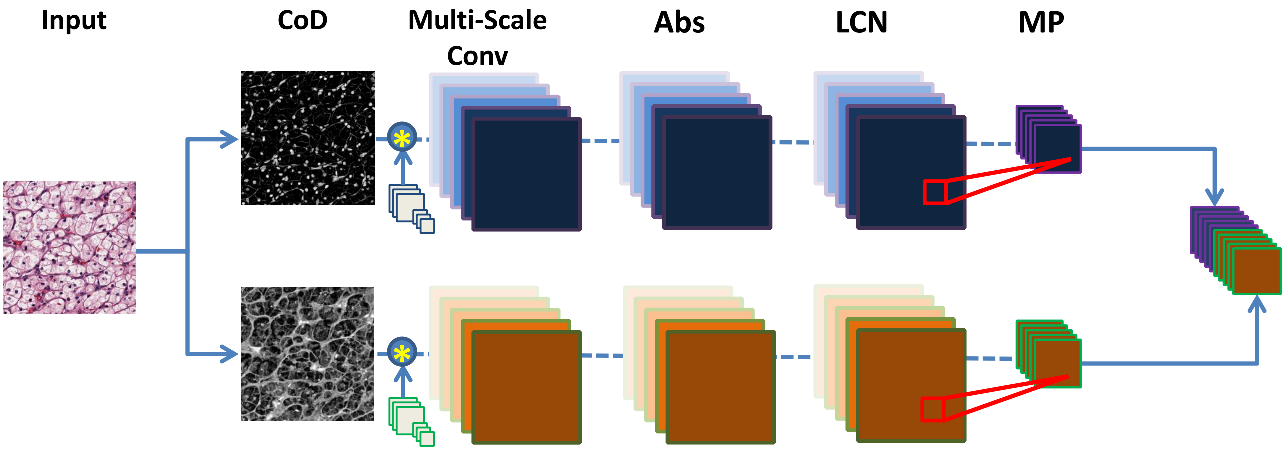


Fig. 2. The proposed multi-scale multi-spectral feature extraction framework. CoD: color decomposition; Abs: absolute value rectification; LCN : local contrast normalization; MP: max-pooling.

feature vectors with dimension SKW. The final representation of the tissue image is constructed as follows,

1) Construct a dictionary $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_P] \in \mathbb{R}^{SKW \times P}$ with P tissue morphometric types, by solving:

$$\min_{\mathbf{B},\mathbf{C}} \sum_{i=1}^{T} \|\mathbf{v}_i - \mathbf{B}\mathbf{c}_i\|^2 \quad (2)$$
$$\text{s.t.} \quad card(\mathbf{c}_i) = 1, \|\mathbf{c}_i\|_1 = 1, \mathbf{c}_i \succeq 0, \forall i$$

where $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_T] \in \mathbb{R}^{P \times T}$ is a set of codes for reconstructing $\mathbf{V}$, cardinality constraint $card(\mathbf{c}_i)$ enforces $\mathbf{c}_i$ to have only one nonzero element, $\mathbf{c}_i \succeq 0$ is a non-negative constraint on all vector elements. Eq. (2) is optimized by alternating between the two variables. After training, the query signal set $\mathbf{V}$ is encoded via Vector Quantization (VQ) based on dictionary $\mathbf{B}$, *i.e.,* assigning each $\mathbf{v}_i$ to its closest tissue morphometric type in $\mathbf{B}$.

2) Construct the spatial histogram for SPM [32]. This is done by dividing an image into increasingly finer subregions and computing local histograms of different tissue morphometric types falling into each of the subregions. The spatial histogram, H, is then formed by concatenating the appropriately weighted histograms of tissue morphometric types at all resolutions, *i.e.,*

$$\begin{aligned} H_0 &= H_0^0 \\ H_l &= (H_l^1, ..., H_l^{4^l}), 1 \le l \le L \\ H &= (\frac{1}{2^L}H_0, \frac{1}{2^L}H_1, ..., \frac{1}{2^{L-l+1}}H_l, ..., \frac{1}{2}H_L) \end{aligned} \quad (3)$$

where $(\cdot)$ is the vector concatenation operator, $l \in \{0, ..., L\}$ is the resolution level of the image pyramid, and H_l represents the concatenation of histograms for all image subregions at pyramid level l. Note, the formulation of spatial histogram, H, is derived from the work in [32], and please refer to Equation 1 and Equation 3 in [32] for details.

For the final classification, the spatial histograms are transformed via homogeneous kernel map [33] for improved scalability with the adoption of linear SVM [34].

4.1.3 Experimental Setup

We have compared the proposed approach with six other approaches on both GBM and KIRC datasets. Implementation details of all approaches are summarized in Table 1. On the implementation of nonlinear kernel SPM , we used the standard K-means clustering for constructing the dictionary and set the level of pyramid to be 3. During evaluation, We repeated all experiments 10 times with random splits of training and test set, and reported the final results as the mean and standard deviation of the classification rates on the following two distinct tumor types:

1) GBM Dataset. It contains 3 classes: Tumor, Necrosis, and Transition to Necrosis, which were curated from whole slide images (WSI) scanned with a 20X objective (0.502 micron/pixel). Examples can be found in Figure 3. The number of images per category are 628, 428 and 324, respectively. Most images are 1000×1000 pixels. In this experiment, we trained on 80 and 160 images per category and tested on the remaining images, with three different dictionary sizes: 256, 512 and 1024. Detailed comparisons are shown in Table 2.
2) KIRC Dataset. It contains 3 classes: Tumor, Normal, and Stromal, which were curated WSI scanned with a 40X objective (0.252 micron/pixel). Examples can be found in Figure 3. The number of images per category are 568, 796 and 784, respectively. Most images are 1000×1000 pixels. In this experiment, we trained on 140 and 280 images per category and tested on the remaining images, with three different dictionary sizes: 256, 512 and 1024. Detailed comparisons are shown in Table 3.

4.1.4 Discussion

1) Joint learning (MSCSC) vs. Separate learning (CSC) for the construction of multi-scale filters. To better understand the difference between joint learning and separate learning in terms of multi-scale filter construction and its impact on classification performance, we designed the comparison between Multi-Scale-CSCSPM and PSEUDO-Multi-Scale-CSCSPM, where the only

Method	Description	
MultiScale-CSCSPM	**Preprocessing**	13×13 Gaussian filter
	Color Decomposition	The input tissue image was decomposed into two spectra corresponding to the nuclear chromatin and the extracellular matrix respectively [26]
	Filter Scale(s)	13×13 and 27×27
	Number of filters (K)	75/150 per spectrum per scale for GBM/KIRC
	Multi-Scale Joint Learning	Yes
	Sparse Regulation Parameter(α)	0.1
	Max-pooling step-size	27
	SPM	non-linear kernel SPM
PSEUDO-MultiScale-CSCSPM	**Preprocessing**	13×13 Gaussian filter
	Color Decomposition	The input tissue image was decomposed into two spectra corresponding to the nuclear chromatin and the extracellular matrix respectively [26]
	Filter Scale(s)	13×13 and 27×27
	Number of filters (K)	75/150 per spectrum per scale for GBM/KIRC
	Multi-Scale Joint Learning	No
	Sparse Regulation Parameter(α)	0.1
	Max-pooling step-size	27
	SPM	non-linear kernel SPM
MCSCSPM [25]	**Preprocessing**	13×13 Gaussian filter
	Color Decomposition	The input tissue image was decomposed into two spectra corresponding to the nuclear chromatin and the extracellular matrix respectively [26]
	Filter Scale(s)	27×27
	Number of filters (K)	150/300 per spectrum for GBM/KIRC
	Multi-Scale Joint Learning	NA
	Sparse Regulation Parameter(α)	0.1
	Max-pooling step-size	27
	SPM	non-linear kernel SPM
PSD^{n}SPM [29]	**Number of Stages (n)**	2
	Patch Size	20×20
	Sparsity	30
	SPM	non-linear kernel SPM
ScSPM [35]	**Patch Size for SIFT**	16×16
	Step Size	8×8
	Sparsity Regulation(λ)	0.15
	SPM	linear SPM
KSPM [32]	**Patch Size for SIFT**	16×16
	Step Size	8×8
	SPM	non-linear kernel SPM
SMLSPM [28]	**Features**	cellular morphometric sparse codes
	Sparsity Regulation(λ)	0.15
	SPM	linear SPM

TABLE 1

Detailed description of experimental setup with different methods. Note the parameters were set empirically to maximize the performance and comparability.

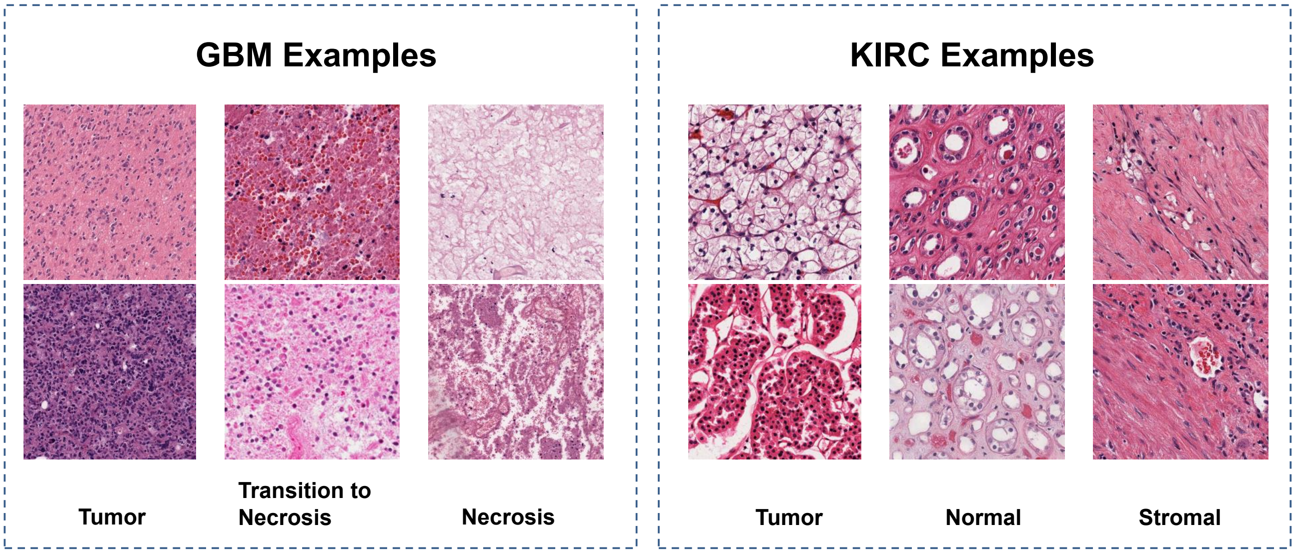


Fig. 3. Examples from GBM and KIRC datasets. Note that the phenotypic signatures are highly diverse in each column.

	Method	DictionarySize=256	DictionarySize=512	DictionarySize=1024
160 training	MultiScale-CSCSPM	93.42 ± 0.61	93.85 ± 0.82	**93.96 ± 0.93**
	PSEUDO-MultiScale-CSCSPM	92.86 ± 0.78	93.11 ± 0.75	93.05 ± 0.71
	MCSCSPM [25]	92.71 ± 0.91	93.01 ± 1.10	92.65 ± 0.75
	PSD^{2}SPM [29]	91.85 ± 1.03	91.86 ± 0.78	92.07 ± 0.65
	SMLSPM [28]	92.35 ± 0.83	92.57 ± 0.91	92.91 ± 0.84
	ScSPM [35]	79.58 ± 0.61	81.29 ± 0.86	82.36 ± 1.10
	KSPM [32]	85.00 ± 0.79	86.47 ± 0.55	86.81 ± 0.45
80 training	MultiScale-CSCSPM	91.50 ± 0.81	92.38 ± 0.87	**92.59 ± 1.01**
	PSEUDO-MultiScale-CSCSPM	91.79 ± 1.18	91.43 ± 1.08	91.26 ± 1.05
	MCSCSPM [25]	91.41 ± 1.07	91.19 ± 0.91	91.13 ± 0.93
	PSD^{2}SPM [29]	90.51 ± 1.06	90.88 ± 0.66	90.51 ± 1.06
	SMLSPM [28]	90.82 ± 1.28	90.29 ± 0.68	91.08 ± 0.69
	ScSPM [35]	77.65 ± 1.43	78.31 ± 1.13	81.00 ± 0.98
	KSPM [32]	83.81 ± 1.22	84.32 ± 0.67	84.49 ± 0.34

TABLE 2
Performance of different methods on the GBM dataset.

difference is that the multi-scale filters in Multi-Scale-CSCSPM were jointly learnt through MSCSC, while the multi-scale filters in PSEUDO-Multi-Scale-CSCSPM were separately learnt at each scale via CSC and concatenated afterwards. Figure 1(a) shows some examples of multi-scale filters jointly/separately learnt from each individual spectrum. It is clear that, through joint learning via MSCSC, the filters at smaller scale (i.e., 13×13) mainly captures lower-level features (e.g., edges), while the filters at larger scale (i.e., 27×27) are more responsible for higher-level features (e.g., complex pattern in extracellular matrix). However, filters learnt separately via CSC do not have such scale-specificity, and present a mixture of both low-level and high-level features at both scales, which might lead to feature redundancy across scales. The difference in scale-specificity becomes more distinct for filters learnt from extracellular matrix, since compared with nuclear chromatin, extracellular matrix sees much more complex patterns, which CSC fails to capture. As a result (shown in Table 2 and Table 3), Multi-Scale-CSCSPM outperforms PSEUDO-Multi-Scale-CSCSPM on both datasets. And we suggest that MSCSC intrinsically allows trainable collaboration of filters across different scales, which potentially leads to filters with improved scale-specificity, and as a result, less feature redundancy across scales.

2) Multi-Scale filters vs. single scale filters. Biological events often express themselves at different scales due to the inherent heterogeneity (e.g., cell type, cell state and the micro-environment). Therefore, the capability to capture and characterize biological events at different scales is very much demanded. For fair comparison, in our experiments, Multi-Scale-CSCSPM, PSEUDO-Multi-Scale-CSCSPM and MCSCSPM were configured

	Method	DictionarySize=256	DictionarySize=512	DictionarySize=1024
280 training	MultiScale-CSCSPM	99.13 ± 0.23	**99.15 ± 0.28**	99.11 ± 0.12
	PSEUDO-MultiScale-CSCSPM	97.08 ± 0.42	97.28 ± 0.40	97.21 ± 0.21
	MCSCSPM [25]	97.39 ± 0.36	97.51 ± 0.41	97.48 ± 0.40
	PSD^{2}SPM [29]	99.03 ± 0.20	98.89 ± 0.19	98.92 ± 0.21
	SMLSPM [28]	98.15 ± 0.46	98.50 ± 0.42	98.21 ± 0.44
	ScSPM [35]	94.52 ± 0.44	96.37 ± 0.45	96.81 ± 0.50
	KSPM [32]	93.55 ± 0.31	93.76 ± 0.27	93.90 ± 0.19
140 training	MultiScale-CSCSPM	98.72 ± 0.50	**98.66 ± 0.39**	98.51 ± 0.68
	PSEUDO-MultiScale-CSCSPM	96.49 ± 0.62	96.67 ± 0.50	96.61 ± 0.49
	MCSCSPM [25]	96.73 ± 0.84	96.89 ± 0.48	96.84 ± 0.67
	PSD^{2}SPM [29]	98.26 ± 0.34	98.07 ± 0.46	97.85 ± 0.56
	SMLSPM [28]	97.40 ± 0.50	97.98 ± 0.35	97.35 ± 0.48
	ScSPM [35]	93.46 ± 0.55	95.68 ± 0.36	96.76 ± 0.63
	KSPM [32]	92.50 ± 1.12	93.06 ± 0.82	93.26 ± 0.68

TABLE 3
Performance of different methods on the KIRC dataset.

to learn the same number of filters, which are 300 and 600 filters for GBM and KIRC datasets, respectively. Experimental results, as summarized in Table 2 and Table 3, show that, for both GBM and KIRC datasets, Multi-Scale-CSCSPM outperforms PSEUDO-Multi-Scale-CSCSPM and yields the best performance. However, PSEUDO-Multi-Scale-CSCSPM only outperforms MCSCSPM on GBM; while becomes less favorable compared to MCSCSPM on KIRC. These observations suggest that,

a) Classification system built on multi-scale filters learnt via MSCSC is more preferable compared to the one built on filters at single scale;
b) Joint multi-scale filter learning (MSCSC) ensures the scale-specificity of filters, and thereafter the consistency of performance for system (i.e., Multi-Scale-CSCSPM) built upon features extracted via such filters.
c) The inconsistency of classification system (PSEUDO-Multi-Scale-CSCSPM), built on PSEUDO-multi-scale filters, attributes to the lack of scale-specificity of such filters, which can potentially leads to feature redundancy across scales, and as a result, less favorable performance even compared with system (i.e., MCSCSPM) built on filters at single scale.

3) Multi-Scale filter learning vs. Multi-Stage filter learning. Compared to the most recently proposed multi-stage unsupervised feature learning system (PSD^{n}SPM) [29], Multi-Scale-CSCSPM consistently achieves better performance over two distinct tumor types, with significantly less number of filters (i.e., 300 vs. 1024 on GBM; and 600 vs. 1024 on KIRC). These advantages, we suggest, are results of i) scale specificity enforced by the proposed multi-scale filter learning strategy; and ii) convolutional filter learning, which, compared to patch-based learning, leads to much more compact filter bank that are translation-invariant.
4) Multi-Scale filter learning v.s. biological meaningful prior knowledge. System built upon biological meaningful prior knowledge (i.e., SMLSPM [28]) can be very effective for the task of tissue histology classification. However, biological meaningful prior knowledge might not always be available straightforwardly (i.e., cellular morphometric properties, as used in SMLSPM, might be difficult to extract), which, as a result, potentially limits the generalization ability of such system to different applications. The results, as shown in Table 2 and Table 3, indicate that the proposed system, Multi-Scale-CSCSPM, is superior to the system built upon biological meaningful prior knowledge (i.e., SMLSPM [28]), without imposing additional requirements (e.g., nuclei are segmentable), which is a better alternative for analyzing large cohorts of distinct tumor types with substantial technical variations and biological heterogeneities.

4.1.5 *Experimental Revisit*

1) Color Decomposition: to investigate the benefit of color decomposition in the proposed tissue histology classification pipeline, we have further evaluated Multi-Scale-CSCSPM with two more variations: Multi-Scale-CSCSPM-RGB and Multi-Scale-CSCSPM-Gray. For Multi-Scale-CSCSPM-RGB, convolutional filter banks were learned from / applied to R, G, and B channels separately, where the number of filters were set to be 50 and 100 per channel per scale for GBM and KIRC, respectively. For Multi-Scale-CSCSPM-Gray, convolutional filter banks were learned from / applied to the grayscale image, where the number of filters were set to be 150 and 300 per scale for GBM and KIRC, respectively. The number of filters were set to ensure the comparability among the variations, and all other experimental setup remains the same as for Multi-Scale-CSCSPM. The best performances on GBM and KIRC

datasets with 160 training images and 280 training images per category, respectively, were illustrated in Figure 4. It is clear that color decomposition is beneficial to tissue histology classification, which is due to the capturing of biological-component-specific information and, therefore, improve the classification performance [25].

2) Max-pooling: to investigate the benefit of max-pooling in the proposed tissue histology classification pipeline, we have further evaluated Multi-Scale-CSCSPM with two more variations: Multi-Scale-CSCSPM-MeanPooling and Multi-Scale-CSCSPM-NoPooling. All other experimental setup remains the same as for Multi-Scale-CSCSPM, and the best performances on GBM and KIRC datasets with 160 training images and 280 training images per category, respectively, were illustrated in Figure 5. It is clear that max-pooling strategy outperforms the other options, probably due to its robustness to local translations.
3) Absolute value rectification: to investigate the benefit of absolute value rectification in the tissue histology classification pipeline, we have further evaluated Multi-Scale-CSCSPM without absolute value rectification: Multi-Scale-CSCSPM-noAbs. All other experimental setup remains the same as for Multi-Scale-CSCSPM, and the best performances on GBM and KIRC datasets with 160 training images and 280 training images per category, respectively, were illustrated in Figure 6. It is clear that absolute value rectification is desirable for the task of tissue histology classification.

4.1.6 Further Comparison with Other Related Work

Existing multi-scale computer vision applications typically concatenate filters from multiple learning layers for multi-scale feature extraction, or use hierarchical pooling to construct the multi-scale features based on single scale filter responses. For the comparison with the former case, Multi-Scale PSD^{2}SPM was implemented with filters concatenated from both the first and the second layers; and for the comparison with the later case, Yang's model [36] was adopted with implementation based on multi-spectral single-scale (27×27) filters, for fair comparison, followed by 3 max-pooling layers in hierarchy (in our experiments, mean-pooling results in $\sim$5% performance drop for both datasets).

Furthermore, it is also very interesting to compare with the Convolutional Neural Networks (CNN) [3], [4] due to its demonstrated success in many different computer vision applications [5]–[7], [37]. Here, we adopted AlexNet [11] and VGGNet [38], which are two of the most successful deep convolutional neural network architectures. During evaluation, we followed the suggestions in [39] with different level of transfer learning settings on both GBM and KIRC datasets using aggressive data augmentation strategies (e.g., flipping, rotation and changing of illumination), among which, we found that the direct application of the pre-train networks (bvlc_alexnet [40] and VGG_ILSVRC_19_layers [38]) produced the best performance. Specifically, for both AlexNet and VGGNet, features were extracted on 224×224 patches with step-size (45) followed by SPM as used for all other approaches.

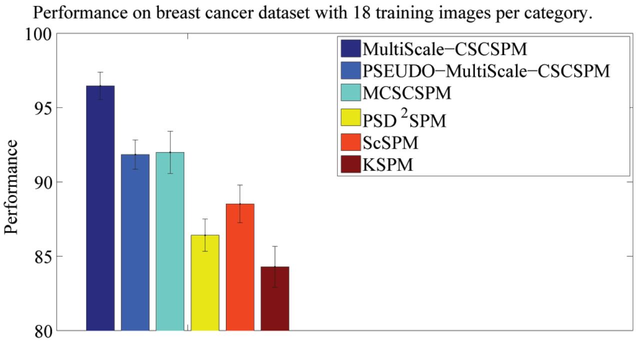


Fig. 9. Performance of different methods for the classification of subtypes in breast cancer.

The best performances on GBM and KIRC datasets were reported with 160 and 280 training images per category, respectively, as shown in Figure 7. It is clear that our proposed method outperforms the pre-trained AlextNet on both GBM and KIRC datasets, while produces highly competitive results compared with the pre-trained VGGNet, where, specifically, it outperforms the pre-trained VGGNet on GBM dataset and underperforms the pre-trained VGGNet on KIRC dataset. The experimental results suggest that the pre-trained deep neural networks from natural domain (e.g., from ImageNet [41]) encode sharable information that is potentially applicable to biomedical domains. Furthermore, given the significant smaller model structure and the unique unsupervised learning capability of the proposed work, compared with deep neural networks, our proposed work provides a highly competitive solution for the learning and application of sharable information with improved computational efficiency and reduced label dependency, which is especially beneficial and desirable to biomedical domains.

4.2 Further Evaluation of MSCSC on Classification of Breast Cancer Subtypes

As a further validation, we have also applied the classification pipeline (MultiScale-CSCSPM) for the classification of subtypes in breast cancer. The dataset for evaluation contains 3 classes: DCIS model, ERBB2+, and Triple Negative, which were collected from 22 breast cancer cell line, and scanned by phase contrast microscope with a 10X objective. Examples can be found in Figure 8. The number of images per category are 36, 40 and 40, respectively. Most images are 1024 × 768 pixels. In this experiment, we trained 18 images per category and tested on the remaining images, with fixed dictionary size: 1024. All experimental protocols and parameter settings were identical to those described in Section 4.1.3, except that no color decomposition was involved (gray-scale image). The final results (see Figure 9) show superior performance of our approach, which confirms the effectiveness and applicability of the proposed multi-scale convolutional sparse coding model to various different tasks.

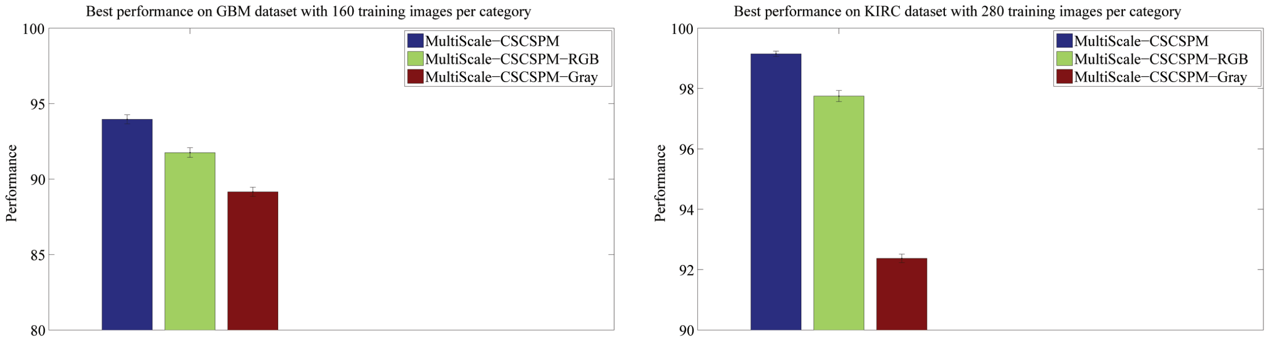


Fig. 4. Experimental revisit on color decomposition, where, by default, MultiScale-CSCSPM operated on decomposed spectra corresponding to the nuclear chromatin and the extracellular matrix respectively.

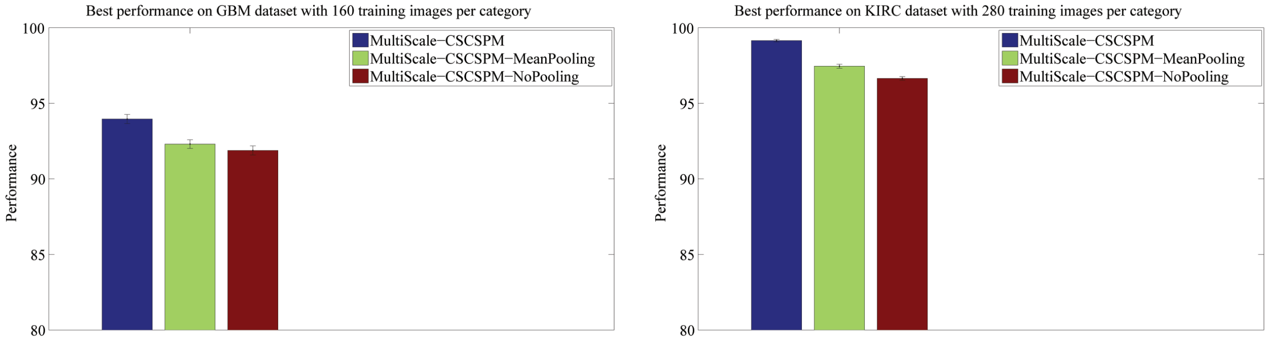


Fig. 5. Experimental revisit on max-pooling, where, by default, MultiScale-CSCSPM utilized the max-pooling strategy.

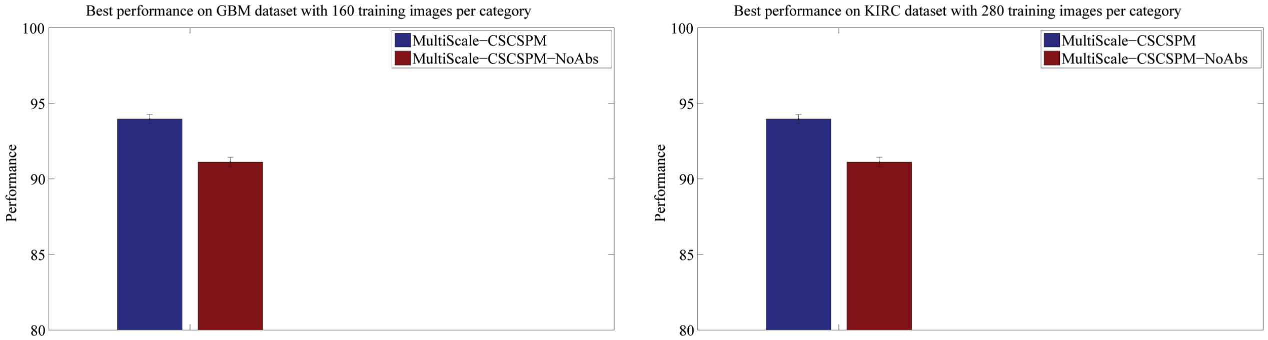


Fig. 6. Experimental revisit on absolute value rectification, where, by default, MultiScale-CSCSPM employed absolute value rectification.

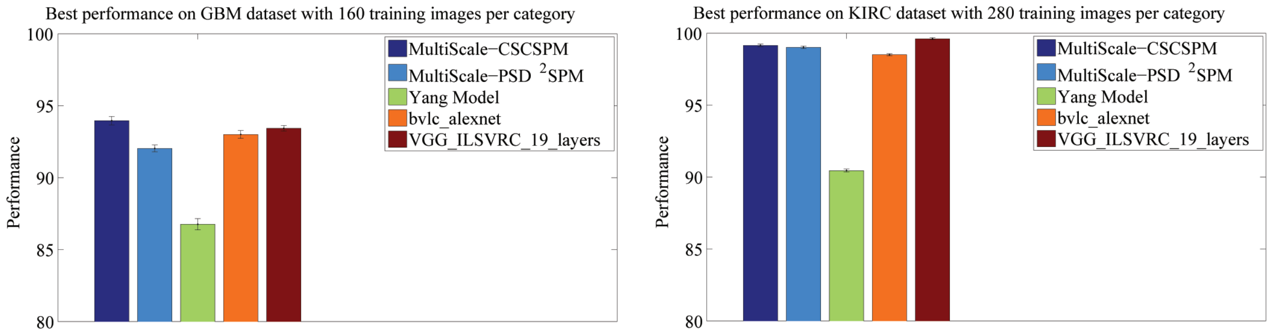


Fig. 7. Comparison with other related multi-scale/deep learning methods, where the best performances of each method/strategy on GBM and KIRC datasets were reported with 160 and 280 training images per category, respectively.

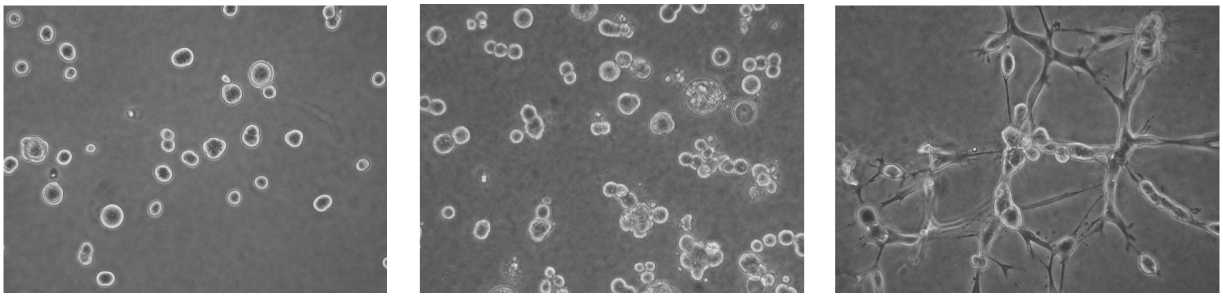


Fig. 8. Examples: First column: DCIS model; Second column: ERBB2+; Third column: Triple Negative.

5 Evaluation of Transfer Learning Capability of MSCSC

The multi-scale joint learning characteristics of MSCSC as well as its capability in capturing scale-specific patterns not only help improve the performances of various regular classification tasks as demonstrated in Section 3, but also provide an unsupervised solution for (I) learning sharable knowledge from a base dataset; and (II) applying/fine-tuning the base knowledge towards the target datasets. This section provides perceptual validation of the sharable knowledge derived by MSCSC across domains, followed by extensive evaluation and discussion on the perceptual insights.

5.1 Perceptual Evaluation

As a perceptual evaluation, we visualized the multi-scale filter banks jointly learned by MSCSC from different domains in Figure 10, which indicates that: (I) filter banks with smaller scale(s) always capture general features regardless of the training domain; (II) the specificity of features captured by filter banks with larger scale(s) trained on various domains, is an increasing function with respect to the dissimilarity among domains; and (III) interestingly, the generality of knowledge, captured by the filter banks with larger scale(s) of MSCSC within different cancer-related domains, still maintains to a large degree, which suggests that the adaptive/transferable knowledge in cancer-related domain(s) is derivable through MSCSC in an unsupervised fashion. These insights are further justified quantitatively as follows.

5.2 Quantitative Evaluation of Sharable Knowledge Across Tumor Types in Tissue Histology from Human Patients

Figure 10 suggests that features learned from histology domain are transferable/sharable across tumor types from human patients. As a quantitative evaluation, we directly applied (I) the pre-trained model from GBM to the classification task in KIRC dataset; and (II) the pre-trained model from KIRC dataset to to the classification task in GBM dataset. All the experimental protocols were identical to the ones described in Section 4.1, and the best performances on GBM and KIRC datasets with 160 training images and 280 training images per category, respectively, were shown in Figure 11.

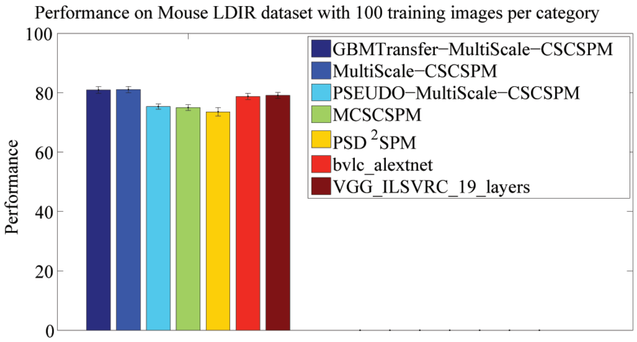


Fig. 12. Evaluation of sharable knowledge derived by MSCSC from human tissue histology for the differentiation of mouse breast tumor morphology between radiation-induced cancer and spontaneous cancer.

5.3 Quantitative Evaluation of Sharable Knowledge from Human to Mouse in Tissue Histology

In this experiment, we were interested to know whether MSCSC trained on human tissue histology can capture sharable information which is applicable to animal models. Specifically, we directly utilized the MSCSC (pre-trained on GBM dataset) for the differentiation of mouse breast tumor morphology between radiation-induced cancer and spontaneous cancer. The dataset contains 2 classes: Sham (control) and LDIR (low dose ionizing radiation at 10 cGy), which were curated from a cohort that was generated for the study of the genetic control of stromal mediation of mammary tumor susceptibility to LDIR [42]. Each category contained 200 images, which were scanned by light microscope with a 40X objective and a fixed size of 2048×1536 pixels. During evaluation, we randomly selected 100 images per category for training and tested on the remaining images with 10 iterations and fixed dictionary size: 1024. All experimental protocols and parameter settings were identical to those described in Section 4.1.3. The final results, as the mean and standard deviation of the classification rates, was illustrated in Figure 12.

5.4 Fine-Tuning Pre-Trained Model from GBM towards Breast Cancer Subtype Classification

With the increase of domain difference, an urgent need is to fine tune the pre-trained model towards new tasks, which can

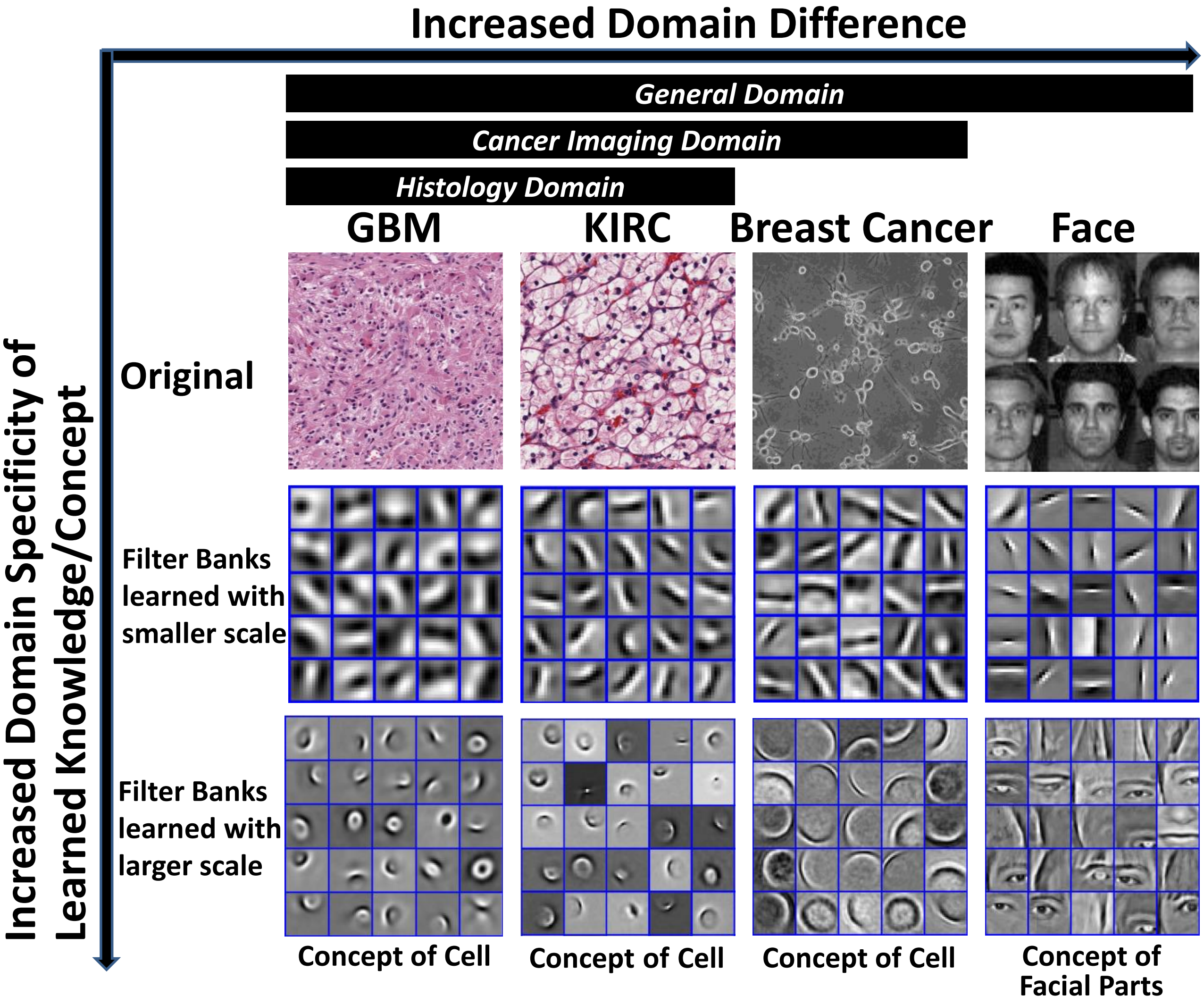

Increased Domain Difference
General Domain
Cancer Imaging Domain
Histology Domain
GBM
KIRC
Breast Cancer
Face
Increased Domain Specificity of Learned Knowledge/Concept
Original
Filter Banks learned with smaller scale
Filter Banks learned with larger scale
Concept of Cell
Concept of Cell
Concept of Cell
Concept of Facial Parts

Fig. 10. Illustration of generality and specificity of features/knowledge derived by MSCSC across domains. It is worth to mention that the filter banks at different scales were jointly learned in an unsupervised fashion with clear scale-specificity: the filters at smaller scale mainly captures lower-level features, while the filters at larger scale are more responsible for higher-level features. **Such an scale-specificity not only help reduce the feature redundancy across scales, but also serves as the basis for transfer learning**.

be easily achieved by fixing the first few filter banks with smaller scales and re-training the rest (filter banks with larger scales) due to the multi-scale joint learning characteristics of MSCSC. As a further justification, we applied the pre-trained model from GBM dataset (see Section 4.1) to the task of Breast Cancer Subtype Classification (see Section 4.2) with different levels of knowledge transfer and tuning settings, and the corresponding performance was illustrated in Figure 13.

5.5 Discussion

Our experimental evaluations above suggest that,

1) The pre-trained multi-scale filter banks by MSCSC may capture sharable knowledge/information across domains, and therefore may be directly applicable to related domain(s). As demonstrated in Figure 10, filter banks at each individual scale all capture similar patterns across different cancer domains, which can serve as the sharable information for tasks across those domains. This insight was further confirmed by the quantitative evaluation of (I) the direct application of pre-trained model from GBM to the tasks in KIRC, Breast Cancer dataset and mouse LDIR dataset, as shown in Figure 11, Figure 13 and Figure 12, respectively; and (II) the direct application of pre-trained model from KIRC to the task in GBM dataset, as shown in Figure 11;
2) The pre-trained multi-scale filter banks by MSCSC can be fine-tuned effectively towards the target dataset in an unsupervised fashion, which is performed by fixing the pre-trained filter bank(s) at smaller scale(s) while re-training the rest (filter bank(s) at larger scale(s)). As shown in Figure 13, the partially-tuned filter banks pre-trained from GBM dataset (GBMTransfer-Multiscale-CSCSPM-ft2nd) saw a steady increase of performance

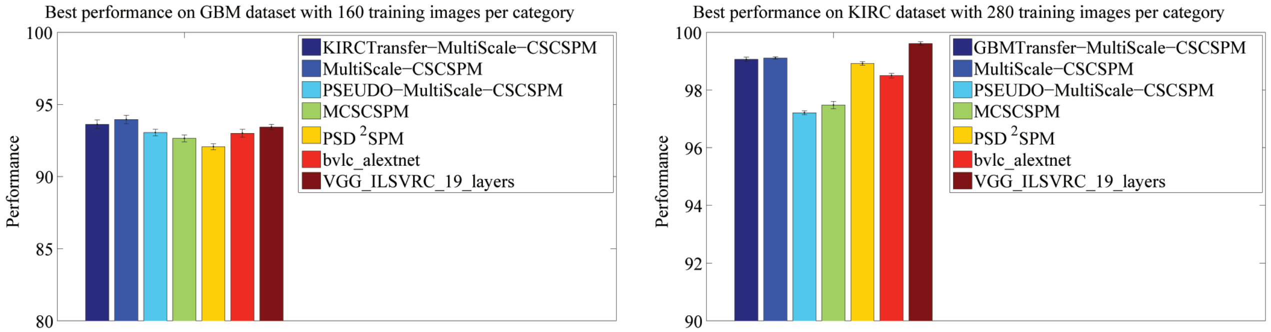


Fig. 11. Evaluation of sharable knowledge derived by MSCSC in the tissue histology domain, where KIRCTransfer-MultiScale-CSCSPM is the direct application of pre-trained model from KIRC to GBM dataset during feature extraction, and GBMTransfer-MultiScale-CSCSPM is the direct application of pre-trained model from GBM to KIRC dataset during feature extraction. It is clear that the information derived by MSCSC independently from GBM and KIRC datasets are directly transferable to each other, which further confirms the insight indicated in Figure 10.

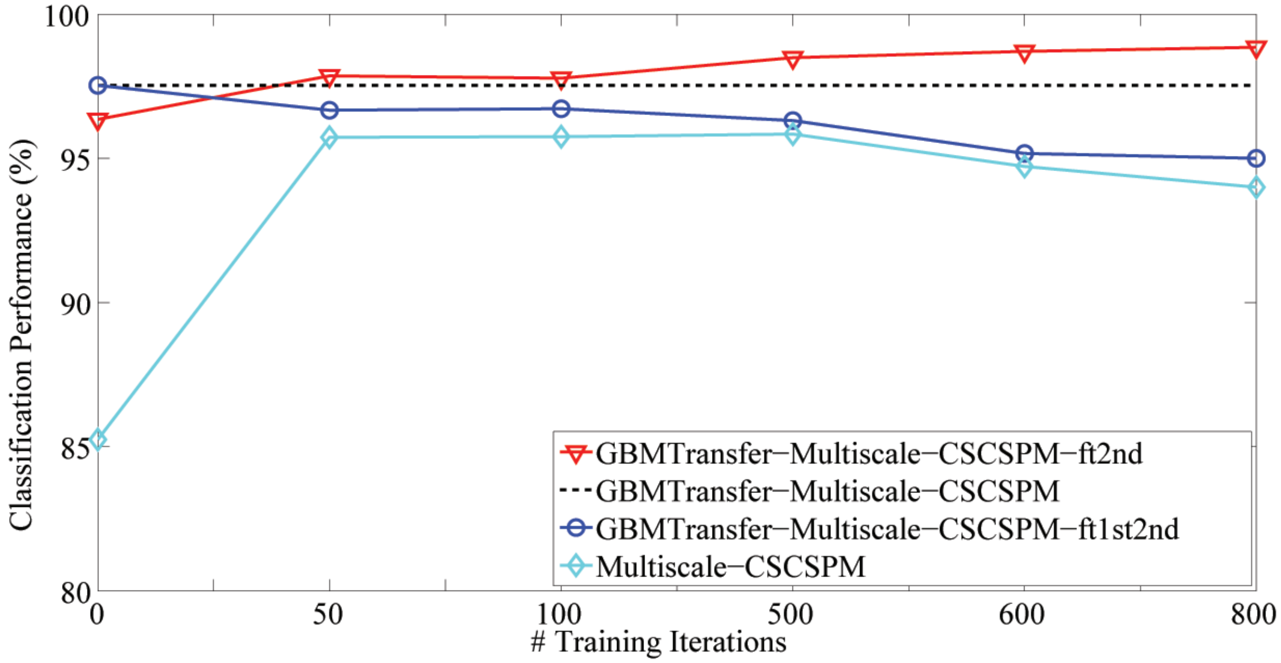


Fig. 13. Experimental evaluation on the transfer learning capability of MSCSC with various transfer/fine-tuning levels, where the filter banks with two different scales were pre-trained on GBM dataset. GBMTransfer-Multiscale-CSCSPM-ft2nd: partial-fine-tuning, where filter bank with smaller scale was fixed and filter bank with larger scale was re-trained/fine-tuned; GBMTransfer-Multiscale-CSCSPM: non-fine-tuning, where pre-trained filter banks at both scales were directly applied without tuning; GBMTransfer-Multiscale-CSCSPM-ft1st2nd: all-fine-tuning, where pre-trained filter banks at both scales were re-trained/fine-tuned; Multiscale-CSCSPM: learning-from-scratch, where filter banks at both scales were directly trained on the breast cancer dataset with random initialization.

on the classification of breast cancer subtypes along filter learning iterations, and both the entirely-tuned filter banks (GBMTransfer-Multiscale-CSCSPM-ft1st2nd) and the filter banks learned directly from breast cancer dataset (Multiscale-CSCSPM) experienced the decrease of performance along filter learning iterations. All the phenomenons are suggested to be tightly related to the scale of target dataset (breast cancer dataset), which, in our case, is significantly smaller compared to the based dataset (GBM dataset).

6 Conclusion

In this paper, we proposed a Multi-Scale Convolutional Sparse Coding model (MSCSC) for unsupervised joint learning of filters at multi-scales with trainable collaboration among them, which, compared to CSC, leads to filters with improved scale-specificity and, subsequently, features with reduced redundancy across scales. Furthermore, such an joint learning strategy also provides an unsupervised solution for transfer learning, which is extremely helpful when the scale of labeled data is very limited. Experimental results, in various biomedical domains, demonstrate the effectiveness of MSCSC on both regular classification tasks as well as its capability in learning sharable base knowledge and fine-tuning it towards new tasks.

Our future work will focus on (I) applying the sharable information learned from GBM/KIRC dataset to a large cohort of tissue histology sections for tumor grading and the association with clinical outcome; and (II) further validating the MSCSC algorithm on various tasks on natural image datasets.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

References

[1] T. S. Lee, D. Mumford, R. Romero, and V. A. Lamme, "The role of the primary visual cortex in higher level vision," *Vision Research*, vol. 38, no. 15/16, pp. 2429–2454, 1998. 1

[2] T. S. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *Journal of the Optical Society of America*, vol. 20, pp. 1434–1448, 2003. 1

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324. 1, 8

[4] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional for generic object categorization," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 2006, pp. 284–291. 1, 8

[5] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*, 2003, pp. 958–962. 1, 8

[6] M. Osadchy, Y. LeCun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Research*, vol. 8, pp. 1197–1215, 2007. 1, 8

[7] B. Kwolek, "Face detection using convolutional neural networks and gabor filters," in *Artificial Neural Networks: Biological Inspirations - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part I*, 2005, pp. 551–556. 1, 8

[8] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006. 1

[9] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, 2009, pp. 609–616. 1

[10] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, 2011, pp. 1017–1025. 1

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 1106–1114. 1, 8

[12] R. Caruana, "Learning many related tasks at the same time with backpropagation," in *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, 1994, pp. 657–664. 1

[13] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, 2012, pp. 17–36. 1

[14] Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski, T. M. Breuel, Y. Chherawala, M. Cissé, M. Côté, D. Erhan, J. Eustache, X. Glorot, X. Muller, S. P. Lebeuf, R. Pascanu, S. Rifai, F. Savard, and G. Sicard, "Deep learners benefit more from out-of-distribution examples," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 2011, pp. 164–172. 1

[15] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 1090–1098. 2, 3

[16] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2528–2535. 2, 3

[17] M. Zeiler, G. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2018–2025. 2, 3

[18] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3626–3633. 2, 3

[19] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua, "Learning separable filters," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 2754–2761. 2

[20] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 391–398. 2, 3

[21] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 647–655. 2

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014, pp. 818–833. 2

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013. 2

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 689–696. 3

[25] Y. Zhou, H. Chang, K. E. Barner, P. T. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 3081–3088. 3, 5, 6, 7, 8

[26] A. Ruifork and D. Johnston, "Quantification of histochemical staining by color decomposition," *Anal Quant Cytol Histology*, vol. 23, no. 4, pp. 291–299, 2001. 3, 5

[27] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 2146–2153. 3

[28] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *Proceedings of the*

Conference on Computer Vision and Pattern Recognition, 2013, pp. 2203–2210. 3, 5, 6, 7

[29] H. Chang, Y. Zhou, P. T. Spellman, and B. Parvin, “Stacked predictive sparse coding for classification of distinct regions in tumor histopathology,” in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013, pp. 169–176. 3, 5, 6, 7

[30] J. Han, Y. Wang, W. Cai, A. Borowsky, B. Parvin, and H. Chang, “Integrative analysis of cellular morphometric context reveals clinically relevant signatures in lower grade glioma,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I*, 2016, pp. 72–80. 3

[31] C. Zhong, J. Han, A. Borowsky, B. Parvin, Y. Wang, and H. Chang, “When machine vision meets histology: A comparative evaluation of model architecture for classification of histology sections,” *Medical Image Analysis*, vol. 35, pp. 530–543, 2017. 3

[32] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178. 4, 5, 6, 7

[33] A. Vedaldi and A. Zisserman, “Efficient additive kernels via explicit feature maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2012. 4

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008. 4

[35] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801. 5, 6, 7

[36] J. Yang, K. Yu, and T. S. Huang, “Supervised translation-invariant sparse coding,” in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, 2010, pp. 3517–3524. 8

[37] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014. 8

[38] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. 8

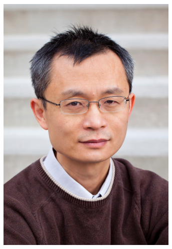
[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 3320–3328. 8

[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the ACM International Conference on Multimedia, MM ’14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 675–678. 8

[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 248–255. 8

[42] P. Zhang, A. Lo, Y. Huang, G. Huang, G. Liang, J. Mott, G. H. Karpen, E. A. Blakely, M. J. Bissell, M. H. Barcellos-Hoff, A. Snijders, and J.-H. Mao, “Identification of genetic loci that control mammary tumor susceptibility through the host microenvironment.” *Sci Rep*, vol. 5, p. 8919, 2015 2015. 10

Hang Chang received his Ph.D. from the Institute of Automation, Chinese Academy of Sciences in 2008. He is a scientist in Biological Systems and Engineering (BSE) Division, Lawrence Berkeley National Laboratory. And the research of his group focuses on knowledge discovery and inference from large scale scientific data with applications to computational biology and biomedical informatics. Current research topics include: (I) Identification of imaging bio-markers towards personalized therapy; and (II) Development of big data oriented open-source Information Technology (IT) solution for domain adaptive cancer informatics.



Ju Han received his Ph.D. degree in the Electrical Engineering Department from the University of California, Riverside in 2005. He is a scientist in Biological Systems and Engineering (BSE) Division, Lawrence Berkeley National Laboratory. His research interests are quantitative and integrative modeling of biological processes.

Cheng Zhong received his Ph.D. from the Institute of Automation, Chinese Academy of Sciences in 2009. He is an affiliate scientist in Biological Systems and Engineering (BSE) Division, Lawrence Berkeley National Laboratory. His research interests are machine learning, computer vision and biometrics.

Antoine M. Snijders received his PhD from the University of Utrecht in The Netherlands. He completed his post-doctoral training at the Cancer Research Institute of the University of California San Francisco. In 2008 he joined the Lawrence Berkeley National Laboratory as a scientist in the Biological Systems and Engineering Division. His lab work focuses on determining the mechanisms that might either predispose or protect an individual from cancer. His laboratory has extensive expertise in using 3D cell culture and mouse model systems to identify phenotypic outcomes associated with environmental exposures by using a strategy that integrates systems genetics and discovery approaches with mechanistic information to ultimately address key questions concerning the effects of environmentally relevant exposures on human health.

Jian-Hua Mao received his PhD at Department of Radiation Oncology, University of Glasgow, UK, completed his post-doctoral training at Department of Medical Oncology, the University of Glasgow. Dr. Mao is a Geneticist Career Staff Scientist in Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory. And he has authored more than 120 peer-reviewed publications. His research interests are: (I) Identify the combinations of genes and their functional polymorphisms that affect the susceptibility to tumor development Discover genetic alterations in tumors using recently developed high throughput technologies, such as CGH microarray, SNP microarray, gene expression microarray, and next generation sequencing; (II) Study the functional and mechanistic role of new discovered genes in tumor development using genetic engineering mice; and (III) Identify the biomarkers for early diagnosis, prognosis and response sensitivity to therapies.