

# Channel Embedding for Informative Protein Identification from Highly Multiplexed Images

Salma Abdel Magid<sup>1</sup>[0000-0001-5791-8128], Won-Dong Jang<sup>1</sup>[0000-0002-5205-1024], Denis Schapiro<sup>2,3</sup>[0000-0002-9391-5722], Donglai Wei<sup>1</sup>[0000-0002-2329-5484], James Tompkin<sup>4</sup>[0000-0003-2218-2899], Peter K. Sorger<sup>5</sup>[0000-0002-3364-1838], and Hanspeter Pfister<sup>1</sup>[0000-0002-3620-2582]

<sup>1</sup> School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

{sabdelmagid, wdjang, donglai, pfister}@g.harvard.edu

<sup>2</sup> Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA

<sup>3</sup> Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

denis\_schapiro@hms.harvard.edu

<sup>4</sup> Department of Computer Science, Brown University, Providence, RI, USA

james\_tompkin@brown.edu

<sup>5</sup> Department of Systems Biology, Harvard Medical School, Boston, MA, USA

peter\_sorger@hms.harvard.edu

**Abstract.** Interest is growing rapidly in using deep learning to classify biomedical images, and interpreting these deep-learned models is necessary for life-critical decisions and scientific discovery. Effective interpretation techniques accelerate biomarker discovery and provide new insights into the etiology, diagnosis, and treatment of disease. Most interpretation techniques aim to discover spatially-salient regions within images, but few techniques consider imagery with multiple channels of information. For instance, highly multiplexed tumor and tissue images have 30-100 channels and require interpretation methods that work across many channels to provide deep molecular insights. We propose a novel channel embedding method that extracts features from each channel. We then use these features to train a classifier for prediction. Using this channel embedding, we apply an interpretation method to rank the most discriminative channels. To validate our approach, we conduct an ablation study on a synthetic dataset. Moreover, we demonstrate that our method aligns with biological findings on highly multiplexed images of breast cancer cells while outperforming baseline pipelines. Code is available at <https://sabdelmagid.github.io/miccai2020-project/>.

**Keywords:** Highly multiplexed imaging · Deep learning · Interpretability.

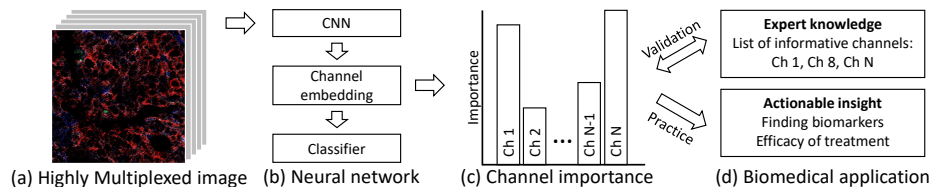


Fig. 1: Informative channel identification. (a) Given highly multiplexed imaging data, we train (b) a neural network to encode a channel embedding and classify a label (*e.g.*, tumor grade). Then, we measure (c) the classification task channel importance by adopting an interpretation method to the channel embedding. (d) We evaluate our system by comparing the predicted informative channels to expert knowledge, and provide new insights for clinicians and pathologists.

## 1 Introduction

Highly multiplexed imaging provides data on the spatial distribution of dozens to hundreds of different protein and protein modifications in a tissue. This provides an unprecedented view into the cells and structures that comprise healthy and diseased tissue. As such, highly multiplexed imaging is emerging as a potentially breakthrough technology in translational research and clinical diagnosis. Examples of highly multiplexed imaging technologies include imaging mass cytometry (IMC) [5], multiplexed ion beam imaging (MIBI) [6], co-detection by indexing (CODEX) [3], and cyclic immunofluorescence (CyCIF) [8].

Each image can comprise 30 to 100 unique channels (that each correspond to the detection of a specific protein) with millions of cells, and so computational tools are essential for analysis. To interpret the outputs of computational tools and answer specific research and clinical questions, it is critical to know which image channels are informative. Even though there is research on interpretation techniques for natural images [14,15,7,18], channel- or target-wise importance ranking interpretation techniques for highly multiplexed images do not yet exist.

We introduce a novel system to automatically identify informative channels in highly multiplexed tissue images and to provide interpretable and potentially actionable insight for research and clinical applications. The process is illustrated in Figure 1. What follows is a description of our system for the goal of identifying the most informative channels for assessing the tumor grade of highly multiplexed images [5]. We first encode each channel using the shared weights of a ResNet18 [4] backbone encoder. To obtain an interpretable representation, which we refer to as channel embedding, we use an embedding encoder. Then, we train a classifier to produce a probabilistic prediction for each tumor grade class. Finally, we measure each channel’s contribution to the tumor grade classification by applying an interpretation technique, Backprop [18], that backpropagates gradients to the channel embedding. In our experimental results, we demonstrate that our system outperforms conventional algorithms [20,4,10] combined with interpretation techniques [15,18] on the informative channel identification task

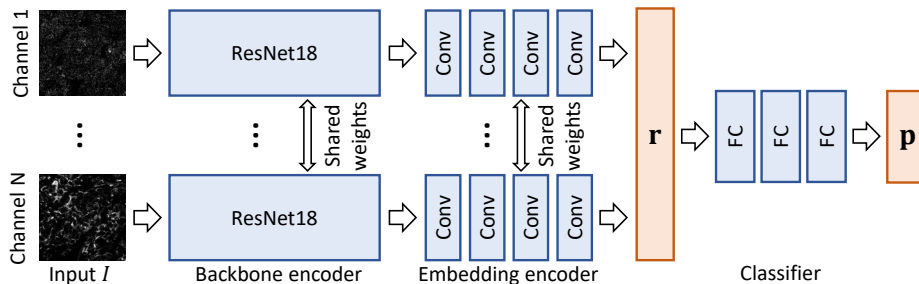


Fig. 2: Our architecture. For an input highly multiplexed image  $I$ , we split channels and feed them into ResNet18 [4]. Next, the embedding encoder extracts an interpretable representation  $\mathbf{r} = [r_0, \dots, r_N]$ , where  $N$  is the number of channels. Both the backbone and embedding encoders share weights across channels. Finally, we adopt three fully-connected layers to estimate class probabilities  $\mathbf{p}$ .

for assessing tumor grade. Moreover, the informative channels identified by our novel method align with findings from a single cell data analysis [5], even though our approach does not require single cell segmentation.

## 2 Related Works

**Interpretation Techniques for Neural Networks:** One category of neural network interpretation is model-agnostic. Backprop [18] and Grad-CAM [15] backpropagate gradients to produce an attention map highlighting important regions in the image. Filter visualization techniques [12,21,1] typically visualize the information extracted from filters. LIME [14], DeepLIFT [17], and SHAP [11] compute the contribution of each feature for a given example. TCAV [7] defines high-level concepts to quantify a model’s prediction sensitivity.

Another category is self-interpretable neural networks. SENN [13] trains a self-explaining model, which consists of classification and explanation branches. Zhang *et al.* [22] modify a traditional convolutional neural network (CNN) by adding masking layers followed by convolution layers to force activations to be localized. Building on this work, Zhang *et al.* [23] visualize a CNN’s decision making process by learning a decision tree for a pre-trained model. However, these methods are not directly applicable to highly multiplexed input images.

**Frame-level Action Localization in Videos:** Discovering informative channels is similar to localizing frames with target actions in videos. Action localization finds frames of interest in an entire video. CDC [16] predicts per-frame confidence scores using 3D convolutional neural networks. BSN [10] and BMN [9] adopt 2D convolutions to estimate actionness, starting time, and ending time at each frame. These methods can be applicable to informative channel identification by using their per-channel classification as a measure of channel importance. However, since they perform prediction by classifying one channel at

a time, their learned features may not be generalizable and consequently their resulting accuracy may not be sufficient.

### 3 Proposed Method

To enable the identification of informative channels, we propose a network architecture with a backbone encoder, an embedding encoder, and a classifier (Fig. 2). First, we will introduce the backbone and embedding encoders, which each extract features from each channel independently. For interpretation, we represent each input image channel as a single value  $r_i$ . Next, we train a classifier that takes the channel embedding  $\mathbf{r}$ , which is a concatenation of  $r_i$  across channels. The classifier yields predictions,  $\mathbf{p}$ . Once our network is trained, we apply Backprop [18] to the channel embedding to produce an attention map, which is then used to rank channels in order of importance.

#### 3.1 Channel Embedding

To leverage knowledge from previously-learned image classification tasks and to extract meaningful features from highly multiplexed images, we begin with ResNet18 [4] pretrained on ImageNet [2] as a backbone network. Naive application of ResNet18 does not allow us to identify channels of interest because it weights information across channels through its constituent convolutions (Fig. 3(a)). Even though 3D CNN [20] can suppress this problem by locally weighting channels, it still blends activations across adjacent channels (Fig 3(b)). Instead, we must extract features from each channel independently. However, doing so will require substantial memory resources (12M parameters per channel). To overcome these problems, we apply the same backbone encoder (shared weights) to each channel individually by modifying the first convolution layer of ResNet18 to accept a single-channel image as opposed to a three-channel (RGB) image (Fig 3(c)). The last fully connected (FC) layer computes a weighted sum of these independently learned channel embeddings thereby leveraging the power of an independent channel encoding technique while still modeling inter-channel interactions.

For interpretation, applying Backprop [18] to the backbone encoder output yields  $N$  2D tensors, where  $N$  is the number of channels. The elements in each 2D tensor represent importance; however, these values are difficult to conceptually interpret. To avoid this issue, we add an embedding encoder that represents each channel as a single value  $r_i$  (Fig 3(d)). The embedding encoder consists of three  $3 \times 3$  and one  $7 \times 7$  convolution layers with batch normalization and ReLU. The  $3 \times 3$  and  $7 \times 7$  convolution layers have 64 and 1 kernels, respectively. Concatenating the embedding encoder outputs yields a channel embedding  $\mathbf{r} = [r_0, \dots, r_N]$ .

To produce a prediction  $\mathbf{p}$  for a classification task, we apply a classifier to the channel embedding  $\mathbf{r}$ . We exploit two FCs with ReLU and one FC with softmax as a classifier. The first two FCs have 200 kernels.

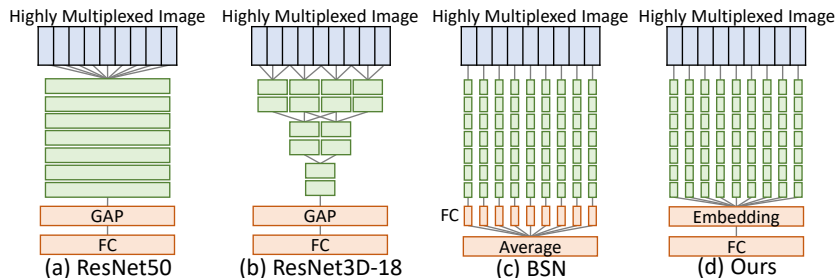


Fig. 3: Architectures of baselines and our network. GAP and FC indicate a global average pooling and a fully connected layer. Unlike other baselines, we convert the input image into a channel embedding and then perform classification.

### 3.2 Informative Channel Identification

After model training, we apply Backprop [18] to the channel embedding. Specifically, we first convert a given highly multiplexed image into a channel embedding and perform classification. We compute gradients at the channel embedding by backpropagating the gradients of the classification output. Then, we set the importance of each channel as the magnitude of its respective gradient. Unlike other systems using standard classification architectures [4,20], our novel system yields a single value at each channel representing its importance in classification. Note that due to this design choice, our system can be used in a plug-and-play fashion with other interpretation methods [14,11,17] as well. We measure the channel importance of all testing images and then average them across images to measure how informative each channel is for classification.

Alternatively, we can apply a linear regression via a single FC layer to the learned channel embedding. Considering that a simple examination of the coefficients is all that is required, this approach may be easier to interpret. We experimented with this design choice and encountered poor results for both datasets. A single FC layer is not sufficient to model the complex channel interactions. To reduce under-fitting, a non-linear combination of the learned embeddings along with Backprop should be used.

**Implementation:** We initialize weights in our network with random values except for the pre-trained ResNet18 backbone network. The spatial resolution of the input image  $I$  is  $224 \times 224$  pixels. For data augmentation, we apply horizontal and vertical flips and random cropping. We use an Adam optimizer with a learning rate of 0.0001. The training process iterates for 100 epochs with early stopping while using a batch size of 32 and four Geforce Titan X GPUs.

## 4 Experimental Results

To validate the design of our model, we conduct an ablation study on a synthetic dataset classification task. To evaluate informative channel identification perfor-

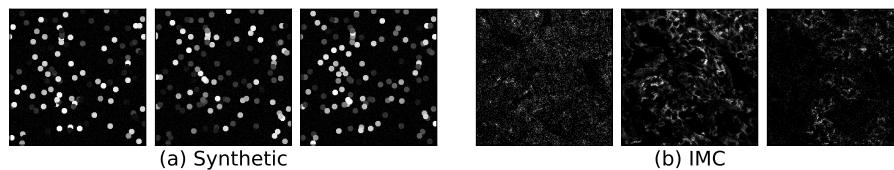


Fig. 4: Sample channels from (a) synthetic and (b) IMC [5] images.

mance, we define a task in which modern deep neural networks [4,20] achieve high accuracy. For a real-world application, we apply our model to a task of predicting tumor grade from a breast cancer dataset generated using IMC [5].

**Methods in Comparison:** To the best of our knowledge, there is no existing method for informative channel identification. As such, we implement baseline pipelines using modern classifiers, ResNet50 [4], ResNet3D-18 [20], and BSN [10], and model-agnostic interpretation methods, Backprop [15] and GradCAM [18]. We apply the interpretation techniques after training the classifiers to compare them with our system. While ResNet50 and ResNet3D-18 directly predict a class from the image, BSN classifies each channel separately and then aggregates the per-channel predictions to generate a single prediction. For ResNet50 and ResNet3D-18, we use Backprop [18] for interpretation. We convert each channel’s attention map into the channel importance by averaging it. For BSN, we adopt GradCAM [15] instead since BSN averages per-channel classification.

#### 4.1 Synthetic Highly Multiplexed Image Classification

According to the single cell analysis of the breast cancer dataset, each channel’s pixel intensities follow a bi-modal distribution. We build a synthetic dataset by emulating this highly multiplexed cell imaging environment. This dataset is then used to conduct an ablation study and validate our design choices. The cellular environment is mimicked in these synthetic images by dispersing a random number of circles (cells) of a fixed radius, whose intensities are sampled from bi-modal distributions (representing the signal arising when a target is either present or absent). We randomly choose two modes between  $[0.1, 0.3]$  and  $[0.7, 0.9]$ . Each mode is randomly assigned a frequency of 0.2 or 0.8. We set the variance of intensities as 0.3. We add Gaussian noise to each image. Since our objective is to identify informative channels, we assign two non-overlapping channels to associate with each ground truth. In particular, the two channels assigned to each class are drawn from a class-specific bi-modal distribution, while the intensities in the remaining channels are drawn from a random distribution. An effective model would identify these specific channels assigned to each ground truth during classification. For example, suppose there are 3 classes, A, B, and C, which are assigned the following channel pairs: 4 and 19, 22 and 31, 2 and 17, respectively. Our model when classifying an image from class C must identify channels 2 and 17 and measure their importance. We synthesize 600 training,

Table 1: Ablation study on the synthetic dataset.

Settings		Scores		
Backbone	Embedding	Accuracy	Recall@6	Recall@10
ResNet18	Independent	98.0	<b>83.3</b>	83.3
ResNet18	Shared	99.7	<b>83.3</b>	<b>100.0</b>
ResNet3D-18	Independent	<b>100.0</b>	50.0	66.7
ResNet3D-18	Shared	99.7	33.3	83.3

Table 2: Quantitative comparison on the synthetic dataset.

Representation	Method	#Param.	Acc.	Recall@6	Recall@10
Single prediction	ResNet50 [4]	24M	<b>100.0</b>	50.0	66.7
	ResNet3D-18 [20]	33M	<b>100.0</b>	50.0	50.0
Channel-wise prediction	BSN [10]	11M	34.7	0.0	0.0
Channel embedding	Ours	12M	99.7	<b>83.3</b>	<b>100.0</b>

300 validation, and 300 test images with 30 channels. Figure 4 visualizes three channels of a synthetic image compared to a real highly multiplexed image [5].

**Evaluation Metric:** To evaluate classification performance, we measure accuracy, which is the number of correct predictions divided by the number of total images. For the assessment of informative channel identification, we use Recall@ $K$ , which is a recall rate when a model proposes  $K$  most informative channels. Since there are six channels (two per class) associated with the classification, we set base  $K$  as 6 and expanded  $K$  as 10.

**Ablation Study:** We conduct an ablation study to find the best architecture design. Namely, we consider two choices for the backbone encoder: ResNet18 [4] and ResNet3D-18 [20]. For channel embedding, we consider two approaches: using shared weights across all channels or using independent channel-wise layers. Table 1 lists the scores of each setting. In terms of classification accuracy, all the settings are comparable. However, the purpose of this experiment is to examine the model’s ability to identify informative channels rather than classification. For informative channel identification, the ResNet18 + Shared setting achieves the best scores in terms of both Recall@6 and Recall@10. Since the backbone encoder, with 3D convolution layers, mixes information across channels, there is performance degradation in the ResNet3D-18 + Independent setting.

**Results:** Table 2 compares our system to the baselines. We find that our novel approach significantly outperforms the others in terms of Recall@6 and Recall@10 while achieving similar classification accuracy. This shows that our channel embedding is highly interpretable and effectively represents each channel.

## 4.2 Tumor Grade Classification

The breast cancer dataset [5] consists of multiplexed images collected using IMC. The images have 39 channels representing a set of proteins that are thought to

Table 3: Quantitative results for tumor grading on the breast cancer dataset [5].

Representation	Method	#Param.	Acc.	Spearman Coeff.
Single prediction	ResNet50 [4]	24M	59.9	23.6
	ResNet3D-18 [20]	33M	<b>68.2</b>	12.9
Channel-wise prediction	BSN [10]	11M	58.5	41.4
Channel embedding	Ours	12M	65.4	<b>61.1</b>

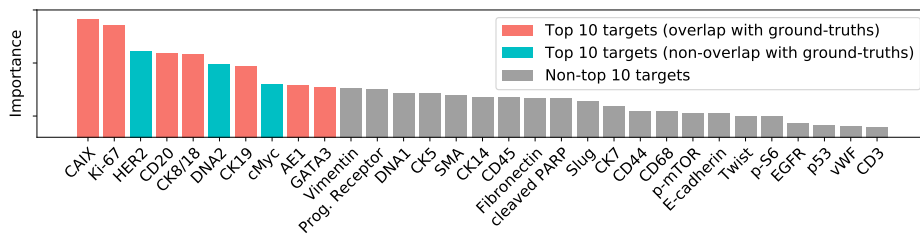
be important for diagnosis or treatment. For each patient, clinical annotation is available. Here, we focus on identifying tumor grade: grade 1, grade 2, or grade 3. Tumor grade is an indicator of disease progression that is typically scored by a pathologist using only H&E images. We seek to identify the most informative channels among the 39 in the dataset with respect to the prediction of tumor grade. The network input is a multiplexed image and the output is a tumor grade. The dataset contains 723 tissue images from 352 breast cancer patients, and we split them into 506 training images and 217 test images. After training our network, we identify informative channels for predicting tumor grade.

**Ground Truth Targets:** Quantification of targets in single-cell analysis is correlated with clinical annotations such as tumor grade; however, this requires segmentation to isolate individual cells [5]. In contrast, we predict the tumor grade without using single-cell segmentation. Additionally, we demonstrate that our approach can interpret the importance of individual proteins (channels). We use the single-cell averaged expression of the individual proteins to compare changes between grade 1, grade 2, and grade 3. Further, we use the sum of the absolute fold-change as a ground truth for analysis. Finally, to evaluate the pipelines, we use the same set of targets as those in the Jackson *et al.* study [5].

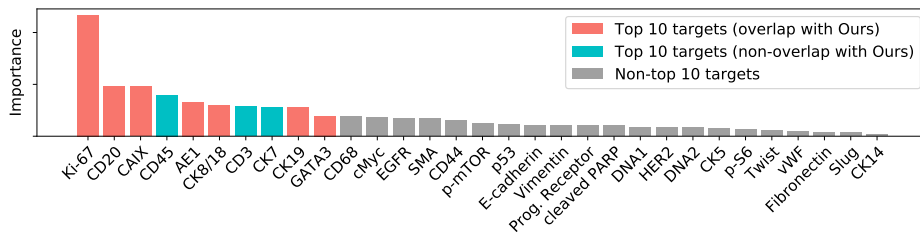
**Evaluation Metric:** For assessment of classification performance, we report the accuracy. To evaluate each model’s informative channel identification, we calculate the Spearman coefficients [19] using the ground truth. The Spearman coefficient measures the correlation between two lists of ranks. To exclude channels with low importance, we only consider the top 15 most informative channels from the ground truth when calculating the Spearman coefficient.

**Results:** Table 3 compares our system to the baseline pipelines. Our classifier performs better than ResNet50 and BSN. For informative channel identification, our system significantly surpasses the baseline systems in Spearman coefficient. Figure 5 shows the importance of each channel predicted by our pipeline compared to the ground truth. We detect seven of the top 10 ground-truth targets, which are known proteins associated with tumor progression. One example is Ki-67, which represents a higher proliferation rate with increasing grade. Another example is CAIX, which represents a marker of hypoxia and also increases with a higher grade. Similarly, it was shown that low cytokeratins (like CK8/18, CK19 and anti-pan keratin (AE1)) are correlated with Grade 3 pathology [5].





(a) Ours



(b) Single-cell derived ground-truth

Fig. 5: Measured target importance for tumor grade classification on the breast cancer dataset [5], ordered by importance. We highlight the top 10 targets.

## 5 Conclusions and Future Work

We have developed a novel pipeline for channel-wise importance interpretation. Our channel embedding effectively simplifies information in each channel while improving channel-wise interpretability. In the experimental results, we show that our pipeline outperforms existing methods [4,10,20] in terms of classification and informative channel identification for tumor grade prediction [5].

Future work will focus on improving the approach and extending it to other datasets and prediction challenges including (i) biomarker discovery associated with survival time in breast cancer [5], (ii) discovery of cellular features predictive of treatment resistance in metastatic melanoma and other diseases, and (iii) the inclusion of spatial transcriptomic data.

## Acknowledgment

Hartland Jackson and Jana Fischer for help with the IMC dataset; This work was supported by NCI (U54-CA225088). Denis Schapiro was supported by an Early Postdoc Mobility fellowship (P2ZHP3.181475). Donglai Wei and Hanspeter Pfister have been partially supported by NSF grant NCS-FO (1835231). Salma Abdel Magid has been partially supported by Harvard University’s Graduate School of Arts and Sciences Graduate Prize Fellowship.

## References

1. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6541–6549 (2017)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., Nolan, G.P.: Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**(4), 968–981 (08 2018)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., Bodenmiller, B.: The single-cell pathology landscape of breast cancer. *Nature* pp. 1–6 (Jan 2020)
6. Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.R., Kurian, A., Van Valen, D., West, R., Bendall, S.C., Angelo, M.: A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174**(6), 1373–1387 (09 2018)
7. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: International Conference on Machine Learning. pp. 2668–2677 (2018)
8. Lin, J.R., Fallahi-Sichani, M., Chen, J.Y., Sorger, P.K.: Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging **8**(4), 251–264 (2016)
9. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: BMN: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3889–3898 (2019)
10. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: BSN: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
11. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
12. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196 (2015)
13. Melis, D.A., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems. pp. 7775–7784 (2018)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. Association for Computing Machinery (Aug 2016)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)

16. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5734–5743 (2017)
17. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3145–3153 (2017)
18. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
19. Spearman, C.: The proof and measurement of association between two things. (1961), publisher: Appleton-Century-Crofts
20. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
21. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)
22. Zhang, Q., Nian Wu, Y., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8827–8836 (2018)
23. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting cnns via decision trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6261–6270 (2019)