Multi-marginal Contrastive Learning for Multi-label Subcellular Protein Localization

Ziyi Liu, Zengmao Wang*, Bo Du*
National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
School of Computer Science, Wuhan University
Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan, China
{ziyiliu, wangzengmao, dubo}@whu.edu.cn

Abstract

Protein subcellular localization(PSL) is an important task to study human cell functions and cancer pathogenesis. It has attracted great attention in the computer vision community. However, the huge size of immune histochemical (IHC) images, the disorganized location distribution in different tissue images and the limited training images are always the challenges for the PSL to learn a strong generalization model with deep learning. In this paper, we propose a deep protein subcellular localization method with multi-marginal contrastive learning to perceive the same PSLs in different tissue images and different PSLs within the same tissue image. In the proposed method, we learn the representation of an IHC image by fusing the global features from the downsampled images and local features from the selected patches with the activation map to tackle the oversize of an IHC image. Then a multi-marginal attention mechanism is proposed to generate contrastive pairs with different margins and improve the discriminative features of PSL patterns effectively. Finally, the ensemble prediction of each IHC image is obtained with different patches. The results on the benchmark datasets show that the proposed method achieves significant improvements for the PSL task.

1. Introduction

The protein subcellular localization(PSL) is essential to interpret and identify the functions of the proteins for revealing the pathology, which can provide valuable information in the target identification process for drug discovery [33, 42]. Analyzing the spatial distributions of human proteins at the subcellular level can help us understand human biology and diseases [14, 27, 32]. For example, the proteins localizing at mitochondria are likely to have the functions of cellular aerobic respiration and energy producing [13]. It also has demonstrated that the abnormalities of subcellular locations of protein are potentially involved in the pathogenesis of many human diseases [8, 22]. Moreover, studying the occurrence of the protein mislocalization under normal and cancer states can help discover and define cancer markers [10].

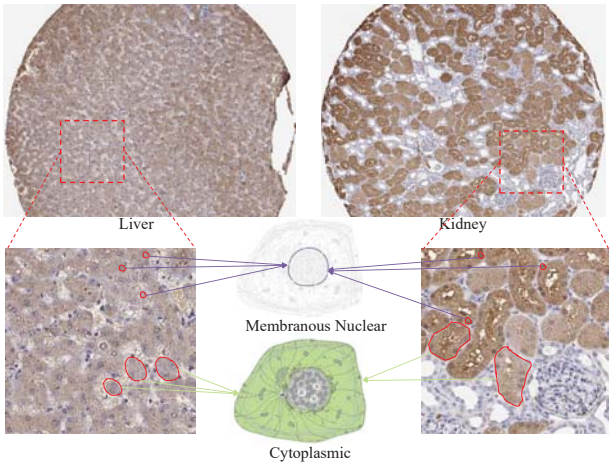


Figure 1. IHC images from liver and kidney show that general structures from two tissues are quite different. The same proteins are displayed on both images using chemical dyes. The PSL labels of both images are nuclear membrane and cytoplasmic. Due to the dyed proteins in cytoplasmic, most of the images are brown. The below detailed image shows that proteins also exist on nuclear membranes.

However, methods that rely on human experts to recognize PSLs, such as wet-lab biological approaches, are time-consuming and expensive. Machine learning is widely used in subcellular pattern recognition to make annotations efficiently. In the past two decades, many studies have focused on PSL in combination with machine learning techniques [21]. According to the data type for PSL, the related researches can be roughly divided into two categories: a) models based on the amino acid sequence and b) models

*Corresponding author.

based on protein high-throughput microscopic images.

Considering that protein functions depend on the amino acid sequence [1, 3, 7], sequence information is obtained to predict protein subcellular locations [15, 19]. However, such methods have low sensitivity in the detection of the dynamic protein translocation, which has been proven to be essential in identifying cancerous biomarkers [2, 4, 16].

Image-based methods usually learn protein distributions with high-throughput microscopic images. Proteins are displayed on images by using chemical dyes or fluorescence, which clearly and concisely reflects protein distributions and spatial expression information [40]. The immunofluorescence (IF) images [24] or immunohistochemistry (IHC) images are two popular images for PSL task. IF images usually need to segment cell cultures or tissue images into single cells, and this task is remarkably challenging [26]. Tissue-based IHC images can show protein distributions from the tissue level to the cell level. Hence, IHC image has been the important source data for PSL task [34]. Compared to the sequence data, IHC images are conducive to studying PSL in healthy and diseased tissues [9].

We show the IHC images from the liver and kidney with the same dyed proteins in Figure 1. From Figure 1, we note that the morphological structures of cells in different tissues are very different despite they having the same PSL. It also should note that the different PSLs in a tissue image may have similar morphological structures. In the light of the fact that nearly 20% of human proteins coexist in more than one subcellular location, many methods are developed for the PSL problem based on multi-label learning [28, 38]. However, the morphological structures with the same PSL cross tissue images and the subcellular differences within a tissue image make it hard to distinguish the distributions of different PSLs, and it is still a challenge to improve the performance of the multi-label PSL methods [26].

In this paper, we propose a new deep learning algorithm, termed the DeePSLoc, to identify protein subcellular locations by using IHC images. To handle the huge size of IHC images, we propose to extract the feature of IHC images with downsampled images and cropped patches. The downsampled image is used to keep the global features of IHC image while the cropped patches are used to keep the details of the structures in the IHC image. Specifically, we use the downsampled image to generate an activation map. Since the activation maps focus on the different morphological structures of the tissues, we select the cropped patches with the highest activation values. Then the global features and the local features can be effectively obtained to capture the morphological and subcellular differences.

To learn the discriminative features of different PSLs, we propose a multi-marginal contrastive learning method in the DeePSLoc architecture, denoted as multi-marginal attention mechanism. The multi-marginal attention mechanism is derived from the self-attention mechanism. We use the contrastive loss with different margins to train such a mechanism. For each margin, we obtain an assignment matrix in which the elements represent how much does the sample pair improve the discriminative ability of features with contrastive learning. With all the positive samples that have the same labels as the anchor sample, we can obtain a positive assignment matrix. Then we generate a positive sample by the weighted average of the original positive samples for contrastive learning and the elements in the assignment matrix are treated as weights. With different margins, we can generate a set of positive samples as well as negative samples for contrastive learning. In fact, these generated contrastive samples have considered the different distributions by learning with different margins. Hence, the contrastive structure effectively reduces the influence of different tissue morphologies whose proteins have the same subcellular locations. The code and models of DeePSLoc are made publicly available at https://github.com/ziniBRC/DeePSLoc. The main contributions of the paper can be summarized as:

- DeePSLoc develops an exciting framework to solve the huge size challenge of IHC images for PSL. It can learn the morphological and subcellular features of IHC images effectively.
- Inspired by the self-attention mechanism, we propose a novel multi-marginal contrastive learning method to generate the contrastive pairs, which can greatly improve the robustness and performance of deep network for PSL. To the best knowledge, this is the first time to weightly aggregate original samples for generating contrastive pairs using attention mechanism.
- The proposed method outperforms the state-of-the-art methods significantly in both single- and multi-label datasets.

The rest of the paper is organized as follows. Related work is discussed in Section 2. Details of the proposed deep learning approach are described in Section 3. The experimental settings and results are presented in Section 4. Conclusions are shown in Section 5.

2. Related Work

2.1. Human Protein Atlas

The Human Protein Atlas(HPA) is a publicly available dataset containing millions of high-resolution IHC images [35]. The IHC images from HPA are brightfield micrographs of two mixed stains that reflect certain protein (brown) and DNA (purplish color) information. Datasets [26] and [39] selected from the HPA usually contain 0–6 images of each protein in dozens of tissues. Figure 1 shows that the effective information of the IHC images from the

HPA. The size of each IHC image is 3000×3000. The composition of datasets brings unavoidable challenges to PSL.

2.2. Tradictional Methods for PSL

Due to the limited classification ability of the traditional classifier, traditional methods often process both DNA channel and protein channel images to integrate more diverse features [18, 25]. iLocator extracted Haralick texture features, DNA distribution features, and LBP features from the separated channels, which characterizes the spatial structure of local image texture and micropatterns [38]. Most PSL algorithms by using IHC images focus on extracting subcellular location features (SLFs) from images [29]. Differing from single-label predictors, multi-label algorithms are used for protein submodular localization by integrating multiple classifiers [28]. The method in [26] attempts to extract the distribution features of protein and DNA in IHC images.

Such methods have some evident limitations. Replacing the original images with the estimated images obtained by the unmixing algorithms will lose image information. Inappropriate selection algorithms may not be able to filter out effective SLFs. The performance of the previous step directly affects the accuracy of the next step. Besides, the framework is not robust enough to the differences in protein distribution across different tissue images.

2.3. Deep Learning Methods for PSL

In recent years, some deep learning based methods have attracted great attention in many fields as well as protein submodular localization with IHC images. AnnoFly [41] leverages CNN to learn the initial feature representations of IHC images and then the RNN network is adopted. The RNN network serves as a classifier by feeding the features from CNN into it. Imploc [23] adopts the pre-trained ResNet model which is trained on ImageNet [17] to extract features from IHC images and then feeds the feature vectors into the transformer network. Due to the huge size of images, all the above networks extract features from the pre-trained backbone networks without fine-tuning. However, there exist domain differences between images from ImageNet and IHC images from HPA. Therefore, features extracted from the pre-trained models which are trained with ImageNet will not adapt to the IHC images classification. Although the performance with deep learning is better than the traditional methods, the methods with satisfactory performance still need to be deeply explored.

3. Methodology

We propose a deep learning model with multi-marginal contrastive learning to predict the protein subcellular localization. In Section 3.1, we introduce the overview of our method. The detailed structures of multi-marginal attention mechanism will be discussed in Section 3.3.

Given the origin IHC image X, we denote the images and cropped patches as X_I and X_P, respectively. We represent the generated positve, anchor and negative samples for image branch and patch branch as $X_I^{\text{g+}}$, X_I^{a}, $X_I^{\text{g-}}$, $X_P^{\text{g+}}$, X_P^{a} and $X_P^{\text{g-}}$, respectively. The model formulas of ResNet backbones in image and patch branch are defined as B_I and B_P, and the ASPP modules [5] in multi-marginal attention mechanism are denoted as F_I and F_P. To make it clear, in this paper, H is the count of the margins, N is the batch size. $\Phi(*)$ measures the Euclidean distance between two samples.

3.1. Overview of the DeePSLoc

The flowchart of the DeePSLoc is shown in Figure 2. Because of the huge size of the IHC images, it's impossible to feed the whole images into deep neural networks without suffering from the out-of-memory problem. We design our model into two branches. Both branches have the same construction with downsampled and patched images inputs. In the first phase, we downsample the original images to low resolution so that we can directly process the IHC images. We aim to train one branch of our network to predict the correct labels with downsampled images. The activation maps are generated according to the output features of backbone networks such as VGG and ResNet, reflecting the discriminative patches for prediction. In the second phase, we crop the discriminative patches with the top T largest activation values. These patches are inputted to another branch of our network to get the local representation. The global features from the downsampled images and the local features from the patched images are concatenated together for the final prediction.

3.2. Data Generation

To avoid out-of-memory(OOM) errors, we downsample the original images to the size of 512×512. Given the trained downsampled images branch, we extract the activation map for each image from the output of the ResNet-18. We calculate the channel-wise average pooling of the features from the backbone. The size of the activation map in DeePSLoc is 16×16. Thus, the original images can be separated into 16×16 sections.

In the patch branch, we crop each original image into several patches into a size of 256×256. When we train our model, one patch serves as one instance for classification, whose label is the same as the uncropped original image. During testing, we crop each testing image into T patches with the size of 256×256 according to the activation map. As shown in Figure 2(d), we average the T predicted probabilities to obtain the final prediction for the testing image. In this way, the challenge of the training network on huge IHC images is relieved, and the prediction that combines the results of patches is enhanced.

For the downsampled images and cropped patches, we

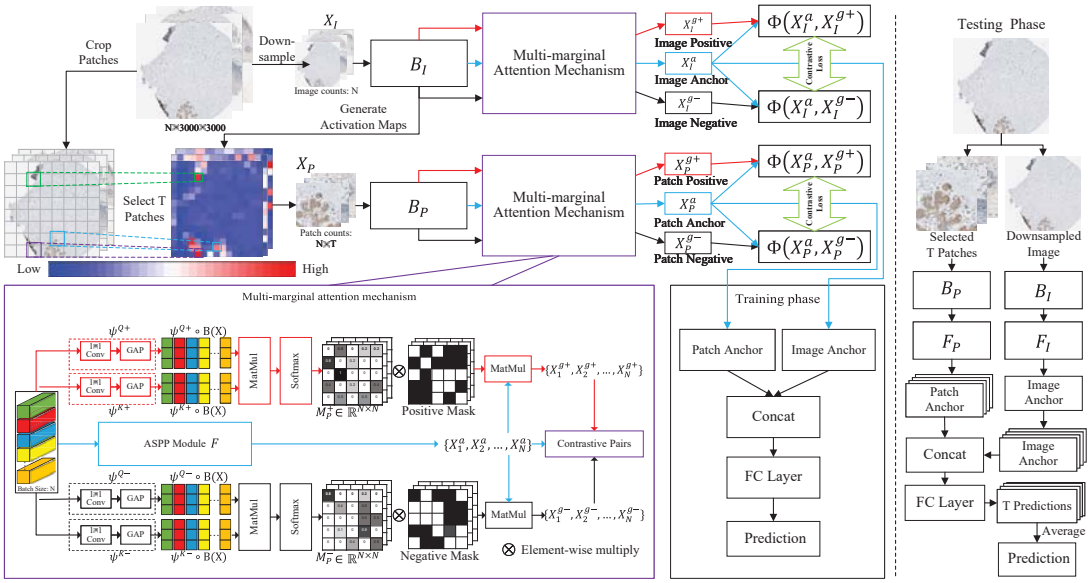


Figure 2. Architecture of DeePSLoc. (a) The original image with huge size is downsampled and cropped into patches. B_I and B_P are backbone networks to extract features from downsampled image and patches respectively. Then the multi-marginal attention mechanism is adopted to construct contrastive pairs of images and patches for contrastive learning. (b) is the detail pipeline of multi-marginal attention mechanism. ψ^{Q+}, ψ^{K+}, ψ^{Q-} and ψ^{K-} represent the 1×1 convolution and global pooling layers. The positive or negative samples are generated by the multiplication of the corresponding assignment matrix and anchor features. (c) shows the classification pipeline during the training phase. (d) The selected T patches and downsampled images are used for prediction.

randomly rotate images between -15 and 15 degrees. Half the images are also randomly horizontally flipped to improve data diversity.

3.3. Multi-marginal Attention Mechanism

Since the distributions of different PSLs across tissue images and within a tissue image have a great difference, the difficulty to align the features between the same PSLs and the different PSLs is also not the same. Hence, it is not reasonable to select anchor samples with the assignment matrix which is obtained with a fixed contrastive margin. To deal with this problem, we train our model to generate easy and hard positive/negative samples for contrastive learning. The target loss for positive/negative samples generation is shown in Figure 3. Lines of different colors denote the loss with different margins. We aim to generate samples that can keep the diversity and distinguishing features of the batch data.

3.3.1 Attention Pipeline

Inspired by the multi-head attention [36], we introduce the attention mechanism into the generation of positive and negative samples with different margins. In multi-head attention, the query, key and value are fed to calculate the attention as:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

For easy understanding, we follow the names of query, key, and value in the multi-marginal attention mechanism.

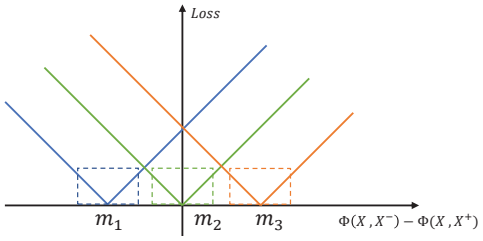


Figure 3. The visualization of multi-marginal loss calculation. The horizontal axis represents the value difference between the $\Phi\left(X, X^-\right)$ and $\Phi\left(X, X^+\right)$. Different colors denote the loss with different margins. For each loss, we aim to train the assignment matrix to learn the pairs around the margin set $\{m_1, m_2, m_3\}$, which are boxed with different colors.

Take the generation of the positive in Figure 2 as an example, the embeddings of the ResNet features are inputted to the attention module, corresponding to the key and query. We compute the dot products of query and key and apply a softmax function to get the assignment matrices. Then, the positive or negative assignment matrix M are calculated as:

$$\begin{aligned} M_i^+ &= \text{softmax}\left(\frac{\psi_i^{Q+}(B_i(X_i))\psi_i^{K+}(B_i(X_i))^T}{\sqrt{d}}\right), \\ M_i^- &= \text{softmax}\left(\frac{\psi_i^{Q-}(B_i(X_i))\psi_i^{K-}(B_i(X_i))^T}{\sqrt{d}}\right), \end{aligned} \qquad (2)$$

where $\psi_i^{\text{K+}}(*)$, $\psi_i^{\text{Q+}}(*)$, $\psi_i^{\text{Q-}}(*)$, and $\psi_i^{\text{Q-}}(*)$ represent the composite functions of 1×1 convolution and GAP layer for

the positive and negative assignment matrces, d denotes the dim of $\psi \circ B(X)$, M_i denotes the assignment matrix from i(image or patch) branch.

3.3.2 Contrastive Pairs Generation

To keep positive/negative samples in the same feature space as anchors, we compute the dot product of positive/negative assignment matrices and anchor features to generate the positive/negative samples. Although the pipelines of positive and negative assignment matrix proceeding are the same, the attention masks of the matrix for further multiplication in Eq. 2 are different. For each anchor, we only consider samples with the same label to generate positives according to the positive assignment matrix. Meanwhile, only samples with different labels are assigned as the negative samples. Thus, we mask the M according to the rules as followed:

$$\begin{aligned} M^{+}_{ij} &= \begin{cases} M^{+}_{ij}, & \text{if } y_i = y_j \\ 0, & \text{if } y_i \neq y_j \end{cases} \\ M^{-}_{ij} &= \begin{cases} 0, & \text{if } y_i = y_j \\ M^{-}_{ij}, & \text{if } y_i \neq y_j \end{cases} \end{aligned} \tag{3}$$

where M^{+}_{ij} and M^{-}_{ij} is the weight that the j th sample is assigned to the ith positive and negative sample, y_i denotes the true label of the ith sample. To make contrastive features the same space, we generate the positives and negatives samples by:

$$\begin{aligned} X^{\text{g+}}_{i,h} &= M^{\text{h+}} F_i \circ B_i(X_i),\ i \in \{I, P\} \\ X^{\text{g-}}_{i,h} &= M^{\text{h-}} F_i \circ B_i(X_i),\ i \in \{I, P\} \end{aligned} \tag{4}$$

where $X^{\text{g+}}_{i,h}$ and $X^{\text{g-}}_{i,h}$ denotes the generate positive or negative samples from i(image or patch) branch by the h^{th} assignment matrix.

3.3.3 Multi-marginal Optimization

For easy understanding, all the formulas below only consider the one branch in our network, X is either from the image branch or patch branch. To train the multi-marginal attention mechanism, the contrastive loss with different margins to train is adopted. it can be represented as:

$$L(\mathbf{m}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=1}^{H} \left| \Phi\left(X^a_i, X^{\text{g-}}_{ih}\right) - \Phi\left(X^a_i, X^{\text{g+}}_{ih}\right) - m_h \right| \tag{5}$$

where X^a_i is the anchor features, $X^{\text{g+}}_{ih}$ and $X^{\text{g-}}_{ih}$ are the generated positive and negative samples from assignment matrix with h^{th} margin m_h, $\Phi(*)$ measures the Euclidean distance between two samples.

We should note that the pairs for contrastive learning are obtained by generating. If we train the proposed architecture with the end-to-end manner, it is hard to be convergent due to that both the inputs and outputs of the multi-marginal attention mechanism are always changing. Hence, in the proposed method, we train the multi-marginal attention mechanism by freezing the backbone networks to guarantee that the features that are used to generate contrastive samples are unchanged. In this way, the architecture of the multi-marginal attention mechanism can be convergent quickly.

3.4. Contrastive Representation Learning

The protein contents and protein distributions on different tissues are different. Given many kinds of tissues, labeling the IHC images in each tissue is expensive. Besides, the available labeled IHC images are usually limited, thereby influencing learning discriminative features.

We define the generated anchor, positive, and negative sample as $(X^a_i, X^{g+}_i, X^{g-}_i), i \in \{1, 2, ..., N\}$, respectively. In this paper, we apply the contrastive structure to the downsampled images and cropped patches.

The contrastive loss can be represented as:

$$L_{con} = \frac{1}{N} \sum_{i=1}^{N} \sum_{h=1}^{H} \max\left(\Phi\left(X^a_i, X^{\text{g+}}_{ih}\right) - \Phi\left(X^a_i, X^{\text{g-}}_{ih}\right) + m, 0\right) \tag{6}$$

where m is a margin, which represents the smallest interval between $\Phi\left(X^a_i, X^{\text{g+}}_{ih}\right)$ and $\Phi\left(X^a_i, X^{\text{g-}}_{ih}\right)$. N is the number of pairs. From Eq. 6, we can observe that the X^a_i is always close to $X^{\text{g+}}_{ih}$ and far away from $X^{\text{g-}}_{ih}$. For easy distinction, we denote the contrastive loss of downsampled images and cropped patches as L^I_{con} and L^p_{con} respectively.

The cross-entropy loss is used to train a classifier effectively and further improve the classification for the single-label prediction task. The cross-entropy loss can be represented as:

$$L_C = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} y_{ij} \log\left(g_j\left(X^a_i\right)\right) \tag{7}$$

We adopt the binary cross-entropy loss for each label in the multi-label scenario. The classification loss can be represented as:

$$\begin{aligned} L_C = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} & y_{ij} \cdot \log\left(g_j\left(X^a_i\right)\right) \\ & + (1 - y_{ij}) \cdot \log\left(1 - g_j\left(X^a_i\right)\right) \end{aligned} \tag{8}$$

where $g_j(X^a_i)$ represents the predicted probability that the i th sample belongs to the jth label, c is the number of labels and y_{ij} denotes the true jth label of the ith sample.

We combine the contrastive loss and the cross-entropy loss together with a tradeoff parameter to learn the representation effectively, and the final loss function can be represented as:

$$L = (1 - \beta) * L_C + \beta * (L^I_{con} + L^p_{con}) \tag{9}$$

where $\beta \in [0, 1]$ is a hyperparameter to balance the importance between classification and contrastive loss.

HPA-7	Acc	Prec	Recall	F1 Score
S.C [26](2008)	66.79	67.31	66.98	66.95
V.C [26](2008)	75.85	76.41	76.05	76.05
ImPLo [23]c(2020)	89.09	89.25	89.09	89.07
Ours	**97.95**	**97.98**	**97.96**	**97.96**

HPA-8	Acc	Prec	Recall	F1 Score
S.C [26](2008)	68.56	71.48	69.84	70.51
V.C [26](2008)	75.49	78.11	76.93	77.43
ImPLoc [23](2020)	84.19	85.06	85.53	85.19
Ours	**95.19**	**96.13**	**95.57**	**95.83**

Table 1. Single-label Classification Results on the HPA-7 and HPA-8 dataset. S.C and V.C denote simple_classifer and voting_classifer [26]. The bold font indicates the best among compared methods.

Multi-HPA	Subset acc	Example acc	Example prec	Example recall	Example F1	Label acc	Label prec	Label recall	Label F1
CSF-CC [28](2018)	89.86	-	-	-	-	-	-	-	-
CSF-BR [28](2018)	85.27	-	-	-	-	-	-	-	-
ML-GCN [6](2019)	85.17	90.23	91.26	92.53	91.89	94.33	89.15	92.61	90.85
ImPLoc [23](2020)	87.93	89.94	90.68	90.80	90.75	94.43	91.83	90.69	91.26
C-Tran [20](2021)	91.38	94.72	95.69	95.92	95.80	97.29	95.89	95.11	95.50
Ours	**95.86**	**96.98**	**97.41**	**97.36**	**97.39**	**98.37**	**97.65**	**96.66**	**97.15**
HPA-18	Subset acc	Example acc	Example prec	Example recall	Example F_1	Label acc	Label prec	Label recall	Label F_1
iLocator [38](2013)	30.3	35.4	40.8	35.6	38.0	77.2	31.1	24.9	27.7
AnnoFly [41](2019)	40.5	44.4	48.8	44.4	46.5	79.9	**91.5**	16.7	28.2
ML-GCN [6](2019)	60.3	**68.0**	74.8	**68.9**	71.5	**89.0**	75.6	36.7	49.3
ImPLoc [23](2020)	53.8	60.8	67.7	61.1	64.2	86.1	81.9	28.3	42.0
C-Tran [20](2021)	57.9	64.9	72.7	64.9	68.6	87.9	86.6	35.0	49.8
Ours	**61.2**	**68.0**	**75.2**	68.5	**71.9**	**89.0**	89.3	**37.1**	**52.4**

Table 2. Multi-label Classification Results on the Multi-HPA and HPA-18 dataset. The bold font indicates the best among compared methods.

4. Experiments

We compare the DeePSLoc with single-label and multi-label methods to evaluate the effectiveness of the proposed method. In each scenario, several state-of-the-art methods are compared, including traditional methods like simple voting classifier [26], CSF classifier [28], and iLocator [38]. For the deep learning methods, we some typical methods for PSL tasks like AnnoFly [41] and ImPloc [23]. Besides, we compare recent deep learning methods using natural images scenes for better clarification, including ML-GCN [6] and C-Tran [20].

4.1. Datasets

To verify the effectiveness of the proposed method, four popular IHC datasets from HPA for PSL tasks are adopted. In the benchmark datasets, there are two single-label datasets and two multi-label datasets in experiments. We choose the HPA-7 [38] and HPA-8 [26] datasets as the benchmark single-label datasets. The Multi-HPA [28] and HPA-18 [23] datasets are used to measure the performance of multi-label methods. More details of the datasets can be found in the Supplementary Materials.

4.2. Implementation Details

In the proposed architecture, we use the widely used network ResNet-18 [11] as the deep backbone network. For each method, we run the experiments 3 times and report the average results. For baseline methods, we set parameters the same as their original papers.

For the parameters in the proposed method, m in Eq. 6 is set as 1, and β in Eq. 9 is set to 0.25. For multiple margins $\mathbf{m}$ in Eq. 3, we set it as the set $\{\pm 1, \pm 0.6, \pm 0.2\}$. With these margins, we can select both the easy and hard pairs for contrastive learning. In multi-label learning, we choose the images that have totally same labels compare to anchors as the positive samples, and the other images as the negative samples. We choose the prediction labels for each image when the probability is larger than 0.5. Meanwhile, in the testing phase, the number T of patches for ensemble prediction is set to 10.

To evaluate the performance of each method, we choose some popular metrics, such as accuracy, precision, recall, and F1 score, for single-label classification task [31]. On multi-label dataset, label-based metrics(accuracy, precision, recall, F1 score) and example-based metrics(subset accuracy, example-based accuracy, precision, recall and F1 score) are adopted as evaluation metrics [23, 37, 43]. The formula definitions of these measures can be found in the Supplementary Materials.

4.3. Results on Single-label dataset

Table 1 shows the classification results of each method on the HPA-7 and the HPA-8 datasets, respectively. Results illustrate that the DeePSLoc can distinguish the protein dis-

tributions on different tissues at the subcellular level by using IHC images effectively. The methods that are trained with SLFs performs much worse than the methods with deep feature on datasets. This demonstrates that it is very necessary to extract the deep features of IHC images for PSL task. In the compared methods, traditional methods like the voting classifier(V. C) perform poorly on HPA-7 and HPA-8, which indicates hows the weakness of traditional methods in extracting features. ImPloc extracts the features of IHC images using pre-trained ResNet18 model. It is pretty hard for ImPloc to present images discriminatively. From Table 1, we can observe that the proposed method outperforms other state-of-the-art methods significantly on almost all the metrics on the HPA-7 and HPA-8 datasets. This demonstrates that compared with these state-of-the-art methods, the proposed DeePSLoc is a promising method for protein subcellular localization with IHC images and it can learn the discriminative features of different PSLs effectively with deep networks.

4.4. Results on Multilabel dataset

Table 2 shows the performance of the methods in the experiments for the multilabel protein subcellular localization task. In the compared methods, the traditional method called the Common-Sets of Features [28] develops two multi-label learning modes: the Binary Relevance (BR) and the Classifier Chain (CC). The CC is better than the BR by considering the correlation between features of different labels. ML-GCN and C-Tran are two state-of-the-art methods by considering the label correlations. In the Multi-HPA dataset, CSF combines the image-level and protein-level features for prediction, while ImPloc and our methods only process IHC images. Although CSF yields better performance than Imploc in CC mode, DeePSLoc outperforms CSF with less information, and the best performance on all metrics is achieved. This demonstrates that the proposed multi-marginal attention mechanism can generate the positive sample and negative sample for discriminative learning of different PSL effectively. In the HPA-18 dataset, all the existing methods including iLocator perform prediction only at the image-level. In general, deep learning methods show great advantages for PSL tasks with IHC images. We should note that DeepSLoc obtains remarkable performance improvement compared to the deep learning methods AnnoFly, ImPloc, ML-GCN, and C-Tran. All the results demonstrate that the proposed DeepSLoc with multi-marginal contrastive learning can improve the discriminativeness of the features for each PSL.

4.5. Ablation Study

4.5.1 Ablation of Contrastive Learning

In Table 3, we show the deep architecture with and without contrastive learning based on different backbone networks,

HPA-7	Acc	Prec	Recall	F1
VGG-11	64.68	66.42	66.16	65.53
VGG-19_bn	74.95	77.17	75.70	75.82
VGG-19_bn+con	83.05	86.60	81.92	83.83
ResNet-18	71.02	73.77	71.94	71.79
ResNet-18+con	**92.44**	**93.42**	**92.82**	**92.88**
ResNet-101	70.75	73.24	71.75	71.38
ResNet-101+con	91.29	94.03	90.76	92.09
ResNet-152	69.20	71.62	68.67	68.15
ResNet-152+con	89.93	93.23	88.66	90.38
DenseNet-121	67.39	72.51	67.19	67.46
DenseNet-121+con	88.32	92.18	88.59	89.77
DenseNet-201	63.15	68.30	63.01	63.16
DenseNet-201+con	87.63	91.86	85.16	87.27

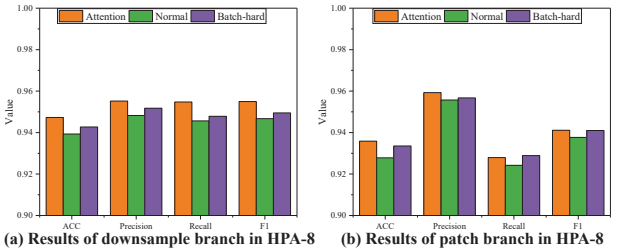
Table 3. Single-label Classification Results of DeePSLoc with only cropped patch input using different backbones on the HPA-7 dataset. The bold font indicates the best among compared methods. The results of model using contrastive learning have been underlined.

such as VGG (11 and 19 bn) [30], ResNet (18, 101, and 152 layers) [11], and DenseNet (121 and 201 layers) [12] to verify the effectiveness of contrastive learning. The classification results on the HPA-7 datasets show that the complexity of a deep model has a remarkable effect on the performance of the subcellular location prediction. The low performance of the VGG-11 indicates that a shallow network cannot learn enough effective features. In addition, the results of the deep DenseNet and the deep ResNet without contrastive loss are not good, and deep networks with high complexity do not improve the model accuracy. When the loss function optimized by the network is combined with the contrastive loss, the performances of all the models are significantly improved. These results can strongly demonstrate that contrastive learning is very effective to learn the discriminative feature of different PSLs cross tissue IHC images by cropping the IHC image into a certain of patches.

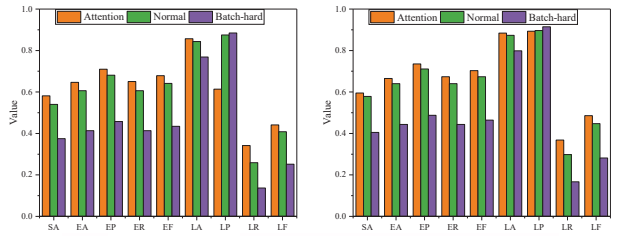
4.5.2 Ablation of Multi-marginal Mechanism

In Figure 4, we attempt to verify the effectiveness of the proposed multi-marginal attention mechanism on HPA-8 and HPA-18 datasets. We compare the proposed method with the methods normal and Batch-hard. Normal is the approach to train the deep architecture with all the triplet pairs while batch-hard trains the deep architecture by selecting the positive samples that are farthest from the anchor sample and the negative samples that are closest to the anchor sample.

From Figure 4, we can observe that the method using multi-marginal attention mechanism, denoted as attention, yields the best performance both in HPA-8 and HPA-18. Compared with Normal, we can observe that Attention has achieved the best performances in almost all the cases. This demonstrates that the mechanism attention mechanism can



(a) Results of downsample branch in HPA-8 (b) Results of patch branch in HPA-8



(c) Results of downsample branch in HPA-18 (d) Results of patch branch in HPA-18

Figure 4. Results of downsample and patch branch in HPA-8 and HPA-18 datasets using multi-marginal attention loss, normal triplets and batch-hard triplets loss. SA, EA, EP, ER, EF, LA, LF, LR, LF represent subset accuracy, example-based accuracy, precision, recall, F1 score, and label-based accuracy, precision, recall, F1 score, respectively.

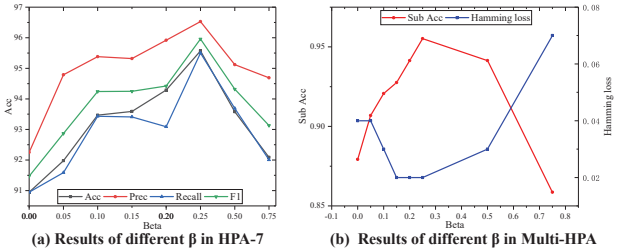
	HPA-7		Acc	Prec	Recall	F1
Single-Label	D		96.18	96.28	96.17	96.19
	P		95.36	95.59	95.29	95.37
	D+P		**97.95**	**97.98**	**97.96**	**97.96**
Multi-Label	HPA-18	Subset acc	Example acc	Example F1	Label acc	Label F1
	D	57.02	63.64	67.15	87.33	43.85
	P	59.50	66.53	67.38	87.33	44.71
	D+P	**61.12**	**68.04**	**71.90**	**88.98**	**52.39**

Table 4. Classification results of our method using downsample images, cropped patches and both on the HPA-7 and HPA-18 dataset. D + P denotes that both downsample images and cropped patches are fed into network. The bold font indicates the best among compared methods.

generate contrastive samples that can improve the discriminative representation ability of different PSLs. Compared with Batch-hard, Attention performs a little poorly on LP. The reason is that Batch-hard predicts all samples with the same labels. This demonstrates that the multi-marginal attention mechanism can improve the generalization ability of deep networks. All these results demonstrate that the multi-marginal attention mechanism is essential in the proposed method.

4.5.3 Ablation of Global and Local Features

Three models are compared in Table 4, denoted as D, P, and D + P, which represent the model only with downsampled images, cropped patches, and both respectively. On the single-label HPA-7 and multi-label HPA-18 datasets,



(a) Results of different β in HPA-7 (b) Results of different β in Multi-HPA

Figure 5. Classification results with different values of β in Eq. 9. The subset accuracy and the hamming loss are metrics for multi-label subcellular localization.

the models that only use downsampled images or cropped patches yield similar performance. When both the downsampled images and cropped patches are used, the proposed method can achieve significant improvements on single-label and multi-label datasets respectively. The results show that the features from downsampled images and cropped patches are complementary, which are both beneficial for classification. More detailed experiment results can be found in the Supplementary Materials.

4.6. Sensitivity Analysis for Parameter β

Figure 5 shows the sensitivity of DeePSLoc in single- and multi-label scenarios with different values of β in single-label and multi-label datasets. Figure 5 clearly shows that When β is larger than 0.25, the performance of DeePSLoc is decreasing quickly. Although contrastive learning is very important, it should be weighted with a proper value for good performance. Hence, we can set β as 0.25 for the practical application.

5. Conclusion

In this paper, DeePSLoc is proposed for protein subcellular localization with IHC images. In DeePSLoc, the global features from downsampled images and the local features from cropped patches are fused for prediction. Activation maps are generated from the downsampled images to select the important patches. The local features are learned effectively and efficiently with these patches. We novelly propose a multi-marginal attention mechanism to softly generate positive and negative samples for better contrastive training at image-level and patch-level, which improve the PSLs across different tissue-based images. Experimental results show that DeePSLoc is promising for PSL.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62006176, 62141112, 41871243, the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170 and the Natural Science Foundation of Hubei Province under Grants 2020CFB241.

References

[1] Yichen Guo A, Ke Yan A, Hao Wu A, and Bin Liu A B. Refold-map: Protein remote homology detection and fold recognition based on features extracted from profiles - sciencedirect. *Analytical Biochemistry*, 611:114013, 2020. 2

[2] Efthalia Angelopoulou, Yam Nath Paudel, and Christina Piperi. Exploring the role of high-mobility group box 1 (hmgb1) protein in the pathogenesis of huntington's disease. *Journal of Molecular Medicine*, 98(3):325–334, 2020. 2

[3] Sebastian Briesemeister, Jorg Rahnenfuhrer, and Oliver Kohlbacher. Going from where to why—interpretable prediction of protein subcellular localization. *Bioinformatics*, 26(9):1232–1238, 2010. 2

[4] Fabian A. Buske, Stefan Maetschke, and Mikael Bodén. It's about time: Signal recognition in staged models of protein translocation. *Pattern Recognition*, 42(4):567–574, 2009. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 6

[7] Kuochen Chou and Hongbin Shen. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mploc 2.0. *PLOS ONE*, 5(4), 2010. 2

[8] Kuo-Chen Chou. Some remarks on predicting multi-label attributes in molecular biosystems. *Molecular Biosystems*, 9(6):1092–1100, 2013. 1

[9] Andreas Digre and Cecilia Lindskog. The human protein atlas—spatial localization of the human proteome in health and disease. *Protein Science*, 30, 2021. 2

[10] Samir M Hanash, Christina S Baik, and Olli Kallioniemi. Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. *Nature Reviews Clinical Oncology*, 8(3):142–150, 2011. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016. 6, 7

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2261–2269, 2017. 7

[13] Mien-Chie Hung and Wolfgang Link. Protein localization in disease and therapy. *Journal of cell science*, 124(20):3381–3392, 2011. 1

[14] Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017. 1

[15] Kenichiro Imai and Kenta Nakai. Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, 10(22):3970–3983, 2010. 2

[16] Darshna M. Joshi, Jignesh Patel, and Hardik Bhatt. In silico study to quantify the effect of exercise on surface glut4 translocation in diabetes management. 2021. 2

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 3

[18] Aparna Kumar, Arvind Rao, Santosh Bhavani, Justin Y. Newberg, and Robert F. Murphy. Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proceedings of the National Academy of Sciences*, 111(51):18249, 2014. 3

[19] Kuo-Chen, Chou, Hong-Bin, and Shen. Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols*, 2008. 2

[20] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 6

[21] Le, Hou, Vu, Nguyen, Ariel, B., Kanevsky, Dimitris, Samaras, and Tahsin and. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern Recognition*, 86:188–200, 2019. 1

[22] KiYoung Lee, Kyunghee Byun, Wonpyo Hong, Han-Yu Chuang, Chan-Gi Pack, Enkhjargal Bayarsaikhan, Sun Ha Paek, Hyosil Kim, Hye Young Shin, Trey Ideker, et al. Proteome-wide discovery of mislocated proteins in cancer. *Genome research*, 23(8):1283–1294, 2013. 1

[23] Wei Long, Yang Yang, and Hong-Bin Shen. Imploc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. *Bioinformatics*, 36(7):2244–2250, 2020. 3, 6

[24] Siyamalan Manivannan, Wenqi Li, Shazia Akbar, Ruixuan Wang, Jianguo Zhang, and Stephen J. Mckenna. An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens. *Pattern Recognition*, 51:12–26, 2016. 2

[25] Murphy, Robert, and F. Building cell models and simulations from microscope images. *Methods A Companion to Methods in Enzymology*, 96:33–39, 2016. 3

[26] Justin Y Newberg and Robert F Murphy. A framework for the automated analysis of subcellular patterns in human protein atlas images. *Journal of Proteome Research*, 7(6):2300–2308, 2008. 2, 3, 6

[27] Katarzyna Radziwon and Amy M. Weeks. Protein engineering for selective proteomics. *Current Opinion in Chemical Biology*, 60:10–19, 2021. 1

[28] Wei Shao, Mingxia Liu, Yingying Xu, Hongbin Shen, and Daoqiang Zhang. An organelle correlation-guided feature selection approach for classifying multi-label subcellular bioimages. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3):828–838, 2018. 2, 3, 6, 7

[29] Aabid Shariff, Joshua Kangas, Luis Pedro Coelho, Shannon Quinn, and Robert F. Murphy. Automated image analysis for

high-content screening and analysis. *Journal of Biomolecular Screening*, 15(7):726–734, 2010. 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. 7

[31] Andong Tan, Duc Tam Nguyen, Maximilian Dax, Matthias Nießner, and Thomas Brox. Explicitly modeled attention maps for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9799–9807, 2021. 6

[32] Peter Thul, Lovisa Akesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Bjork, Lisa M Breckels, et al. A subcellular map of the human proteome. *Science*, 356(6340), 2017. 1

[33] Md. S. Uddin, Abdullah Al Mamun, Md. Ataur Rahman, Tapan Behl, Asma Perveen, Abdul Hafeez, May N. Bin-Jumah, Mohamed M. Abdel-Daim, and Ghulam Md Ashraf. Emerging proof of protein misfolding and interactions in multifactorial alzheimer's disease. *Current Topics in Medicinal Chemistry*, 2020. 1

[34] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220), 2015. 2

[35] Mathias Oksvold Uhlen, Per Fagerberg, Linn Lundberg, Emma Jonasson, Kalle Forsberg, Mattias Zwahlen, Martin Kampf, Caroline Wester, Kenneth Hober, and Sophia. Towards a knowledge-based human protein atlas. *Nature Biotechnology*, pages 1248–1250, 2010. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[37] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020. 6

[38] Yingying Xu, Fan Yang, Yang Zhang, and Hongbin Shen. An image-based multi-label human protein subcellular localization predictor (ilocator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, 29(16):2032–2040, 2013. 2, 3, 6

[39] Yingying Xu, Fan Yang, Yang Zhang, and Hongbin Shen. Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning. *Bioinformatics*, 31(7):1111–1119, 2015. 2

[40] Yingying Xu, Lixiu Yao, and Hongbin Shen. Bioimage-based protein subcellular location prediction: a comprehensive review. *Frontiers of Computer Science in China*, 12(1):26–39, 2018. 2

[41] Yang Yang, Zhou Mingyu, Fang Qingwei, and Shen Hong-Bin. Annofly: annotating drosophila embryonic images based on an attention-enhanced rnn model. *Bioinformatics*, (16):2834–2842, 2019. 3, 6

[42] Yun Zeng, Rachel L. Nixon, Wenyan Liu, and Risheng Wang. The applications of functionalized dna nanostructures in bioimaging and cancer therapy. *Biomaterials*, 268:120560, 2021. 1

[43] Fengtao Zhou, Sheng Huang, and Yun Xing. Deep semantic dictionary learning for multi-label image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3572–3580, 2021. 6