Self-supervised semantic segmentation based on self-attention

Vaidehi Pandey
University Of Twente
Enschede, Netherlands
v.o.p.pandey@student.utwente.nl

***Abstract*—The unavailability of a significant number of annotated images is affecting the medical imaging area of computer vision. This unavailability is because of the fact that labelling a medical image is a time consuming task as it requires careful analysis of the whole image and can only be done by an expert. Supervised learning method will not give good results due to lack of input-output pairs. Self-supervised learning gives us a way out by transforming the images themselves to labels thus not requiring human-annotated labels in larger amounts.**

This paper proposes a self-supervised strategy for representation learning which can be further used for other downstream tasks. In our method we integrate colorization task into BYOL which is a contrastive learning method. The resulting self-supervised method is trained on cem500k dataset with two different encoders namely resnet50 and stand alone self-attention.

The encoders trained through our self-supervised training method achieved comparable results to the encoders trained with the original BYOL. Further, the self-attention model pre-trained using our method performed better than the rest of the encoders on the semantic segmentation task. We analyzed the Class Activation Map(CAM) and found that the self-attention encoder(pre-trained using our method) activates visually important regions on the image.

***Index Terms*—Computer Vision, Semantic Segmentation, Self-supervised Learning, Self-attention, Medical Imaging**

I. Introduction

Semantic segmentation is one of the most complicated but at the same time important task of computer vision. It is challenging because of the pixel level operation where a label is assigned to each pixel belonging to a particular class. It is important because of its application in autonomous driving [4], categorizing clothing items [7], efficient agriculture [18] and in the medical domain for organ segmentation [31] or cellular level segmentation [2].

Deep convolution networks have achieved great results in semantic segmentation by using supervised learning [4] [17] [30] [24]. The networks proposed in [4] [17] [30] are trained on natural images data sets which contain large amounts of high quality accurately labelled images. These images can be easily collected and labelling does not require any expertise. In the case of medical images, firstly acquiring a large number of images is impossible and secondly only an expert can label these images. And labelling medical images is a time consuming task. In the paper [2], authors reported that experts spent 32-36 hours annotating a microscopy data-set consisting of 165 images.

Self-supervised learning is a type of unsupervised learning where data itself provides supervision. This supervision is provided by devising a task called pre-text task that the network solves on a large unlabelled data set in a supervised manner. The network with the learned representation is then fine-tuned on a small labelled data for downstream tasks like segmentation, classification etc. This kind of training method has gained popularity and has achieved results [11] [12] comparable to the supervised state-of-the-art approach. Although a self-supervised learning method requires a lot of data to learn the features, the data need not be labelled. And due to this self-supervised learning method can be used to learn representation from these unlabelled images.

In this research, a self-supervised pre-training method to learn representation from unlabelled microscopy images has been proposed. First, we perform a analysis of various self-supervised methods. Next, we propose a new method and devise the following question to test our new method: **Can a self-supervised learning method with a self-attention model learn a robust representation for semantic segmentation?** The research question has been divided into the following sub-question:

R1): What are the different methodologies in the field of semantic segmentation?
- a) What are the different methods in semantic segmentation?
- b) What are the different self-supervised methods developed in the computer vision domain?
- c) What are the different methodologies developed for self-supervised semantic segmentation?

R2): How does a model with pre-trained BYOL features compare against traditional supervised convolution network?
- a) How does the encoders derived from incorporating colorization task in BYOL pre-training method perform on semantic segmentation benchmark datasets?
- b) What difference in performance does the byol pre-trained encoder have on the semantic segmentation task?
- c) To what extent does the inclusion of colorization task in BYOL pre-training affect the ability of the encoder to learn relevant features?
- d) What does the self-attention model see in an image to make a prediction?

We answer the research question by reviewing the relevant work done in Section II. In addition to this we also show the motivation for our method in section III. In section IV we discuss the proposed method in detail followed by dataset description and experiment setup in section V. Next, we discuss the results obtained in our experiment in section VI and finally conclude our findings in section VII.

II. Related Work

A. Supervised Learning

Supervised learning is a training method that learns a function for mapping an input to an output. The training is carried out on well annotated data meaning that for every input data that the function takes there is already an output. This makes the learning process more accurate. In case of image segmentation, learning to predict accurate masks is of utmost importance and that is the reason why models such as FCN [17], U-Net [24], DeepLab [4] trained using supervised learning methods have gained popularity. All these models have an Encoder-Decoder architecture where the encoder network extracts features from the input image and feeds it into the decoder to reconstruct the segmentation mask.

U-Net: This architecture is the most popular supervised segmentation technique which was developed on medical data sets but has been extended to other domains as well [25]. It has a symmetric architecture of encoder and decoder with skip connections between them. The encoder is usually a pre-trained deep neural network like resnet or VGG which learns the features. The decoder consists of up-sampling and concatenation of features learned by the encoder part. The skip connection helps to combine low and high level features together. The end result of this architecture is a dense representation. The paper [24] reports that the proposed method achieves state-of-the-art results on three different data sets.

Attention-UNet: Skip-connections used in up-sampling in traditional U-net [24] provide spatial information but the regions highlighted in the input image by it can be imprecise. The authors of [21], proposed a method where they applied an attention layer to the skip connections to provide only the important region of interest. The addition of an attention layer leads to better performance but at the expense of more computation time and parameters.

Since the conception of traditional U-net, there have been many variants developed and applied outside the medical domain [1]. Other than U-Net, deeplab [4] is a very popular network for semantic segmentation that has been trained for scene understanding in autonomous driving.

B. Self-supervised Learning

Self-supervised learning leverages the data itself to make predictions without the need of human annotated labels. Since there are no labels available for the model to learn from, self-supervised learning consists of pretext downstream tasks.

Pretext tasks are pre-designed tasks for networks to solve and in process learn visual features about the image. These tasks are designed on the image data itself (usually unlabelled data).

Downstream tasks such as segmentation, classification are computer vision applications that are used to evaluate the quality of features learned by self-supervised learning [14].

Pretext task is the main component in self-supervised learning and the choice of task determines the performance of the learned model on the downstream task. Some of the popular tasks are colorization [29] [16], image inpainting [22], jigsaw puzzle [20] [15], image generation [10] [32].

Colorization as a pre-text task involves predicting the color version of the image given the grayscale image. [29]in his work converts the original image into Lab space and then given the L space of the image, the model is trained to predict the ab space. The authors have defined this as a multinomial classification problem with class re-balancing to distribute the predicted ab values evenly.

Another important pre-text task is Context Prediction. In [22], the authors proposed a method to predict image context by training a CNN to in-paint a missing patch in the image. They reported that if the patch was created in the center of the image then the network would learn about the features in the center. They proposed to create random patches for the network to learn better features. But this random creation of patches results in intensity change. The resultant image belongs to a different domain and the features learned may not be useful.

The pretext task of solving the Jigsaw Puzzle consists of predicting the correct order of jumbled patches of an image. The authors of [19] proposed a method where a random crop in the image was selected. This cropped image is divided into 9 equal shaped patches and then permuted as per pre-defined order. The task of the deep neural network is to predict the order of permutation and in the process learn features. The performance of the proposed method solely depends on the order in which the shuffle is performed and it can be an expensive process to find the perfect shuffle order.

C. Contrastive Learning

It is a machine learning technique used to learn representation such that similar images stay together and dissimilar images are far apart. It is used in both supervised and unsupervised approaches. It is more useful to use this methodology in the self-supervised setting in the absence of labels. SimCLR, MoCo, BYOL are some popular contrastive learning methods.

SimCLR: SimCLR is an acronym for Simple framework for Contrastive Learning. It learns visual representation by maximizing loss between dissimilar images (negative pair) and minimizing loss between similar images(positive pair). In the paper [6], authors sampled N images from the dataset and applied two augmentations to each of the N images. After this augmentation each image has 2 positive images(called positive pair) because two different augmentations were applied. And each of the positive pairs has $2(N-1)$ negative images. Each image in a positive pair is passed through an encoder to get image representations. The encoder used in the paper is ResNet50

but it can be any convolution architecture. This representation is passed through a series of non linear transformations to get an embedding vector. Thus for each augmented image an embedding vector is retrieved. Contrastive loss is calculated using the calculated embedding vectors minimizing distance between positive images and maximizing distance between the negative images. This method relies on negative samples and the batch size has to be large to have enough negative samples for the network to learn.

BYOL: The authors of [11] points out that [6] is sensitive to the choice of data augmentation and mentions that if color distortion is removed the method does not perform as well as it does with its inclusion. This can also bring systematic bias to the model. They proposed a method that discards dissimilar images and only relies on two networks that interact with each other to learn representation. This strategy makes the training process efficient and removes the possibility of systematic bias. The paper reports that this method achieves 74% when pre-trained on ImageNet [8] using ResNet50 as a backbone model.

The architecture of the model is shown in fig 1. The method of BYOL is very simple. First we take one image, apply augmentations or a set of augmentations on the image to get two different views of the same image. These two views are then passed through two different identical network architectures separately, ResNet50 is the backbone as mentioned in the paper. The output of both the networks is passed through a Multi Layer Perceptron(MLP). The output of the online network(upper network in the Fig1) is again passed through a MLP and the output is compared with the second network's (lower network in Fig1) MLP output using L2 loss. The weights of the online network are updated by gradient descent and that of the target network through the exponential moving average(EMA) of the online network.

D. Attention

Attention mechanism in computer vision is divided into two parts: hard and soft attention. Hard attention focuses on a subset of an image, and this makes this approach compute and memory efficient. But one downside of this approach is that these models are difficult to train as the image cropped or sliced (operation that achieve hard attention) may or may not contain relevant features. Soft attention on the other hand acts upon the whole image and is hence memory and computation expensive but the model can be trained using back-propagation. Soft attention has been widely used in computer vision tasks like image classification [28], recognition and segmentation [21] because of its ability to capture long range contextual relationships. Next, we describe different soft attention models.

1) Self-Attention: It is a type of self-attention where the attention is computed by taking the image itself as input. Mathematically, given an input tensor from a previous layer of shape (H,W,F) where H, W and F are the height, width and number of input filters. The tensor is then flattened to a matrix $\mathbf{X} \in \mathbb{R}^{HW \times F}$. The formula to compute the self-attention is:

$$o_h = softmax(\frac{(XW_q)(XW_k)^T}{\sqrt{{d_k}^h}})(XW_v)$$

where W_q, W_k $\in \mathrm{R}^{F \times \frac{dh}{k}}$ are query and key weight matrices and W_v $\in \mathrm{R}^{F \times d_v h}$ is value matrix. o_h is the attention score per head. d_v and d_k denotes the depth of the value and key, query. And ${d_k}^h$ denotes the depth of key for each attention head. XW_q, XW_k and XW_v gives the query Q, key K and value V.

2) Criss-cross Attention: The conventional self-attention [27] establishes contextual information of an image but this information calculated is both memory and computation expensive. The self-attention has complexity of O(N^2), where N denotes the number of pixels in an image. Criss-Cross attention [13] module on the other hand calculates the feature map by considering only the horizontal and vertical positions for a given position. This strategy reduces the complexity to O($\sqrt{N}$) and the memory consumption is also reduced. This method has achieved state of the art results for semantic and instance segmentation tasks on Cityscapes, ADE20K, COCO, LIP and CamVid.

3) Stand Alone Self-attention: : Convolution operation consists of multiplying a filter of fixed size (3x3, 5x5 etc) with each and every position of an input tensor. Convolution layers are translation equivariant but lack the property of capturing long range interaction. Global attention [27] which captures these long range dependencies are computationally expensive. The authors of this paper replaced the convolution layer in the neural network with local self-attention which can be applied to both small and large images. The proposed attention layer performs computation on a small neighbourhood called the memory block. And for each memory block, a single headed attention is computed. The input feature is divided depth-wise into N groups, attention is computed on each group and concatenated. On top of this, positional information is added by computing the relative distance of (i, j) to every pixel. The row and column distances are computed separately and later multiplied with the query matrix.

This approach reduces the number of parameters and the attention is computed for a small block instead of the whole input tensor. The paper compares two types of network architecture, wherein the first model has all the convolution layers replaced by attention and the second where convolution at the top of the layer is preserved and rest is replaced by attention. The latter performed better owing to the fact that the convolution layer better captures the local features.

E. Transformer Model

The Transformer [27] model became popular after its usage in the NLP (Natural Language Processing) domain for solving machine translation tasks and later used in other NLP tasks. The architecture of a transformer consists of encoder and decoder modules and each of these modules further consists of several encoder and decoder models. Each encoder and decoder consists of self-attention and feed forward networks. Feed forward network consists of non linear activation applied on a linear transformation. The training process of the transformer model consists of a pre-training step followed by fine tuning of the model on the downstream task. Pre-training

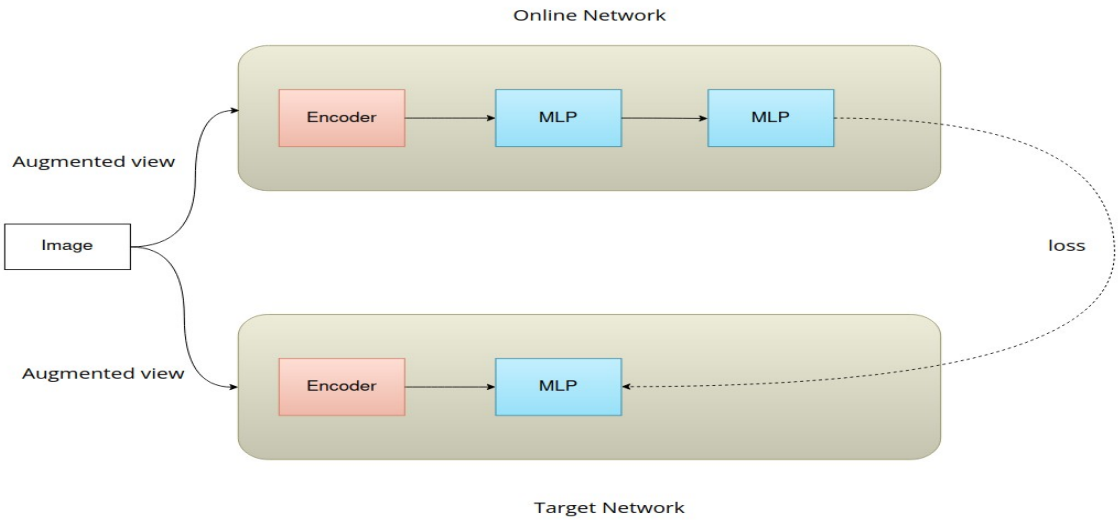


Fig. 1: BYOL Architecture: BYOL consists of two identical networks which takes in 2 different augmented views of the same image. The upper part of the network is the online network and the lower side of the network is the target network. The output of the encoder in the online network is passed to 2 MLPs and that of the target encoder is passed through 1 MLP. The output of the online network after the 2 MLPs and that of target network after 1 MLP are compared with $L2$ loss function. The weights of the online network are updated through gradient descent and that of target network are updated as exponential moving average of the online network.

is done on a large data set in an unsupervised way and the weights learned are fine-tuned on a small data set.

1) Vision Transformer: It is acronym-ed as ViT [9] is the first application of transformers in the computer vision domain. The architecture of the model is similar to the one in [27] with the difference in the encoder's input. The 2D image is divided into patches and each patch is flattened, a positional embedding is added and fed into the encoder. The positional embedding helps to retain the positional information about the sequence of the flattened patch. The encoder consists of alternating layers of self-attention and feed forward network. This model is first pre-trained on larger data and later fine-tuned on Imagenet. One downside of the transformer is its lack of vision related inductive bias. And it is due to this shortcoming that the pre-training phase requires larger data.

Transformer models achieve good results but at the expense of computation cost [5] and large training data. Transformers lack inductive bias, hence requires large amounts of data to train and this can be a shortcoming that is very difficult to subdue.

F. Self-Supervised Semantic Segmentation

In the method proposed in [22], a patch in the image is created for the model to predict and learn important features. But this patch created changes the intensity of the image. The resultant image belongs to a domain different from that of the original image. The authors of [3] proposed a novel method to overcome shortcoming of [22]. The method selects random two isolated small patches in a given image and swap their context. Repeat these operations a number of times, till the intensity distribution is still preserved, but its spatial information is altered. The model consists of two parts namely analysis and reconstruction. In the analysis part the features are learned from the disordered image and later these learned features are used to reconstruct images. One problem with this method is that it uses L2 loss which blurs the image.

In [26], the authors used image inpainting as a pretext task. The network consists of coach and in-painting networks that compete against each other. Coach network learns to create difficult patches for the in-painting model to predict and the in-painting model in return learns features that it uses in the reconstruction of the image. The paper proposed to use ResNet-18 instead of AlexNet, removed the bottleneck layer and used a pre-trained encoder decoder. Coach Network on the other hand also used ResNet-18 to learn features. The loss functions for the inpainting model consists of a reconstruction loss and a context loss. The loss function of the coach network is adversarial to the reconstruction loss function thus creating a competition between the two networks. Their model performed better than popular self-supervised methods on potsdam, SpaceNet and DG Roads datasets.

Also, the representation learned by byol method is later used for semantic segmentation on cityscape data set and this method of pre-training performs better than the pre-training method of Simclr [6] and moco [12].

III. Motivation

We have made certain design choice and here we motivate these choices in detail.

A. Medical Image Scarcity

Acquisition of large amounts of medical images is a difficult task as compared to natural images. And even if acquired

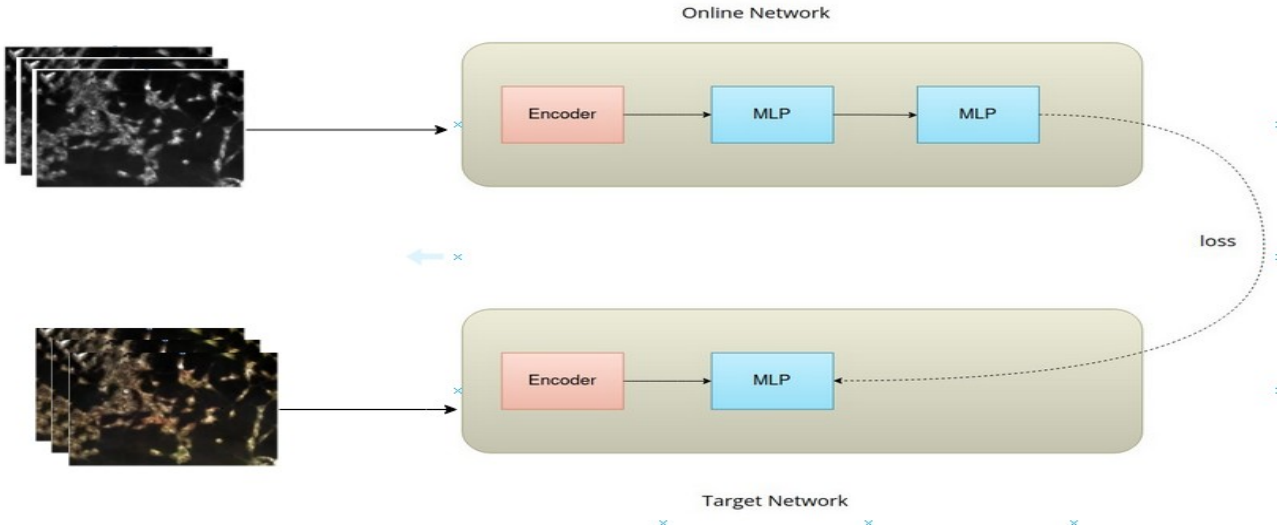


Fig. 2: Modified BYOL Architecture: The proposed method replaces the augmentation with the colorization task and keeps the entire BYOL architecture intact. That means the loss is computed by comparing the representation of online and target network which takes grayscale and colored image as input respectively. Similar to the original BYOL method, another loss is computed by interchanging the inputs to the online and target network. The addition of these two losses are backpropagated to the online network. And the wieghts of the target network are the exponential average of the online.

in large numbers labelling the images is a time consuming process as it can only be done by an expert. Hence, to overcome this issue we will have opted to use a self-supervised approach [11] [29] thus requiring no labels.

B. Colorization as a pretext task

Colorization as a pretext task: Colorization task will allow the model to learn local features of the image. By local features we mean the cellular structures present in the image. This assumption is based on the fact that in an image a particular artifact will have the same color, so the model will learn about the pixels that constitute the artifact group them together and will apply the same color.

C. BYOL for self-supervised learning

The performance of BYOL is superior to other contrastive methods and it also does not require negative samples (Section II). Microscopy images have cellular structures of different sizes and we want to leverage BYOL method of representation learning to learn those features.

D. Stand Alone Self-Attention as Encoder

The main function of Self-Attention is to highlight the prominent features of an image. Our data set consists of cellular level images and the shape of the cell varies with each image and each image can have numerous cellular structures. The use of Self-Attention will give more importance to the cellular structure. Also, the self-attention method in [23] is computationally less expensive and can be applied to the whole image without down-sampling. Traditional Convolution networks lack the ability to long-range interaction and its fixed filter size makes it a template matching task and does not capture semantic features. Also, the global self-attention [27] can only be used after the image has been sampled down and is computationally expensive. But the stand alone Self-Attention is computationally less expensive than self-attention [27] and has less number of parameters than ResNet50.

IV. Method

In this section we discuss the proposed self-supervised learning method and further elaborate on the downstream tasks devised to evaluate the performance of self-supervised strategy.

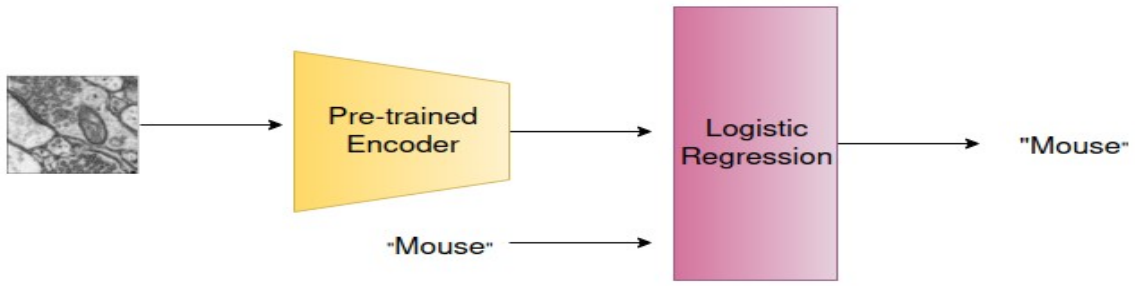
A. Self-supervised training

Fig. 2 shows the architecture of our self-supervised method. Our method consists of the following components.

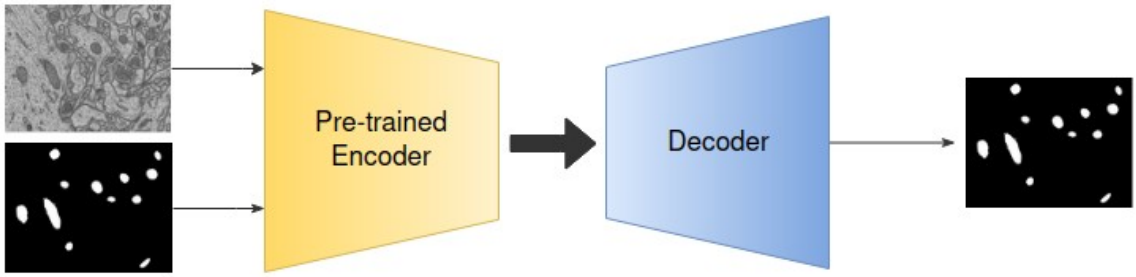
- **Colorization:** The colorization task converts the single-channel grayscale image to a three-channel Lab image. The pre-trained model proposed in [29] has been used to convert the image (cem500k dataset) from grayscale space to Lab space. The gray-scale and the colorized image forms an input-target pair for the encoders as shown in fig. 6.

- **Encoders:** The choice of encoders is an important aspect of this method. In this experiment, Resnet50 and stand-alone self-attention has been used to learn representation. While resnet50 is a well-known resnet network, stand-alone self-attention is a modification of traditional resnet50 where the convolution layer is replaced by the attention layer except in the first layer.

- **BYOL:** The architecture of our proposed self-supervised learning method is shown in fig 2. The original BYOL

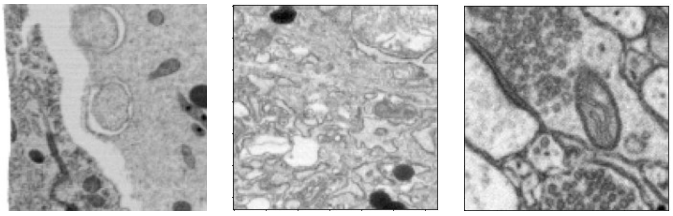


(a) Classification Model



(b) Semantic Segmentation Model

Fig. 3: Downstream Tasks: The above figure shows the architecture of the downstream tasks. The representation learned through our self-supervised learning strategy is evaluated using these two tasks. (a) The pre-trained encoder takes an image (cem500k dataset) as input and outputs a representation which acts as input to the logistic regression classifier. The classifier is now trained in a supervised manner to predict the class of the organisms the cellular image belongs. (b) U-net is the base model used to design the semantic segmentation architecture. The encoders are pre-trained using our method and the decoder is randomly initialized. Lucchi++ and Kasthuri++ are used to perform semantic segmentation. During the training, only the weights of the decoder are updated.



(a) C elegans. (b) Human. (c) Mouse.

Fig. 4: Classification Dataset: The figure depicts example of images belonging to the subset of cem500k data-set used for the classification task. Each image bears the name of the organism to which it belongs.

architecture has been preserved which consists of two identical encoder networks(online and the target).

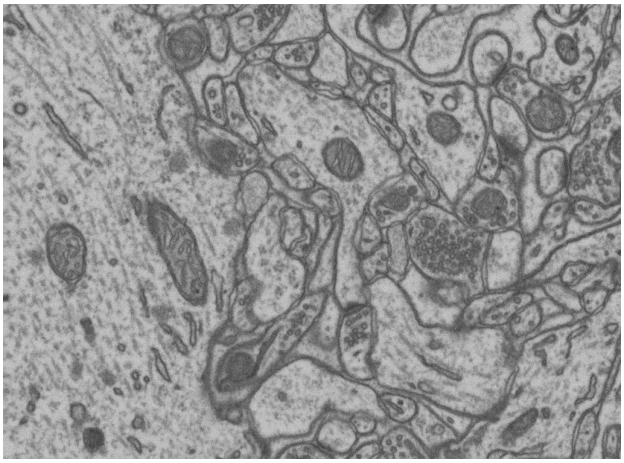
The online network learns from the representation of the target network. That means, the online network takes an image as input and the target network takes the augmented view of the same image as input. In our work, the augmented view has been replaced with the colorized image and the original image is gray-scale which is the input to the online network.

The weights of the online network is updated by back-propagating the loss(L2 loss) which is calculated by comparing the representation of the online and target network. And the weights of the target network are updated as the exponential weighted average of the online network.

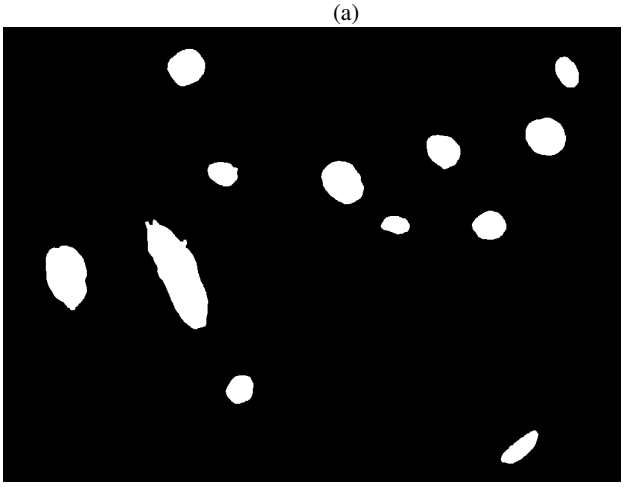
B. Downstream Tasks

The evaluation of the self-supervised method takes place by using the representation learned during pre-training to perform downstream tasks like classification and segmentation. In our experiment we have chosen to perform classification and semantic segmentation.

- **Classification:** The classification task is performed by training a linear classifier on top of a pre-trained encoder. Fig 3a shows the architecture of the classifier used in our experiment. The pre-trained encoder takes in input an image and the output of the encoder which is of size 2048 is used as input for the logistic regression model. Other than this, the regression model also takes the target value in the form of 'name of the organism' (Humans, mouse, c.elegans) to which the image belongs. The logistic regression model is trained to predict the type of organisms(the name of the organisms to which the cellular images belong.). The images used for training

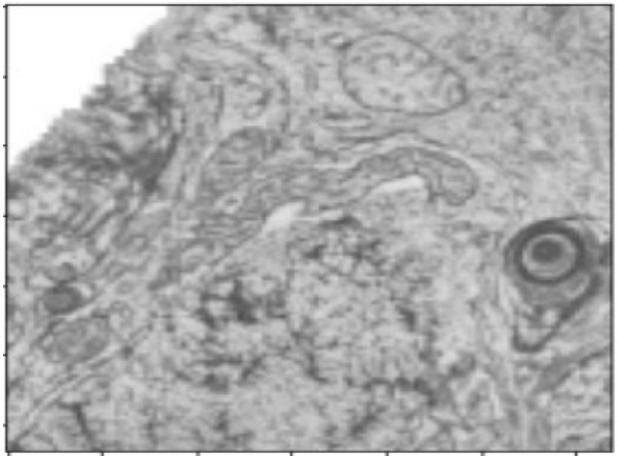


(a)

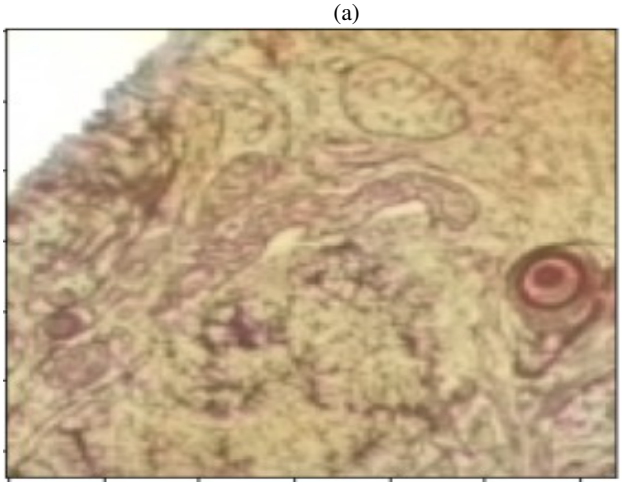


(b)

Fig. 5: Semantic Segmentation sample: The above figure shows the image and its segmentation mask of Lucchi++ data set. Fig 5a shows the captured cellular level image of the brain cells of a mouse. Fig 5b shows the segmentation mask of the cellular image.



(a)



(b)

Fig. 6: Pre-training dataset sample: The images present in the cem500k dataset are gray-scaled asshown in (a). The self-supervised method proposed in this paper requires grayscale and color image pair. The pre-trained network proposed in the paper [29] has been used to convert the grayscale to color image as shown in (b)

the classifier belong to a subset of the cem500k data set. Its distribution is shown in Fig. 7a and Fig. 4 shows 3 images belonging to that subset used in this downstream task.

- **Semantic Segmentation:** Semantic segmentation is the task of grouping together similar parts of the image that belong to the same class. The architecture of the segmentation model is shown in Fig. 3b. It consists of two parts namely the encoder and decoder. The encoder takes in the image and its corresponding mask as input. The output of the encoder acts as input for the decoder and eventually the decoder predicts the segmentation mask. The encoder in the architecture is pre-trained in a self-supervised manner. During the training, the weights of the encoder remain unchanged and only the decoder's weights are updated by back-propagation. This method allows us to evaluate the representation learned during pre-training. U-net architecture has been chosen as the base model for this task. Resnet-50 and stand alone self-attention are the encoders for two different U-nets.

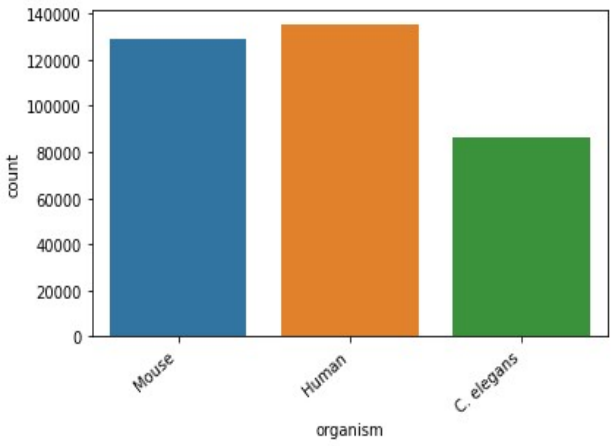
V. Experiments

A. Dataset

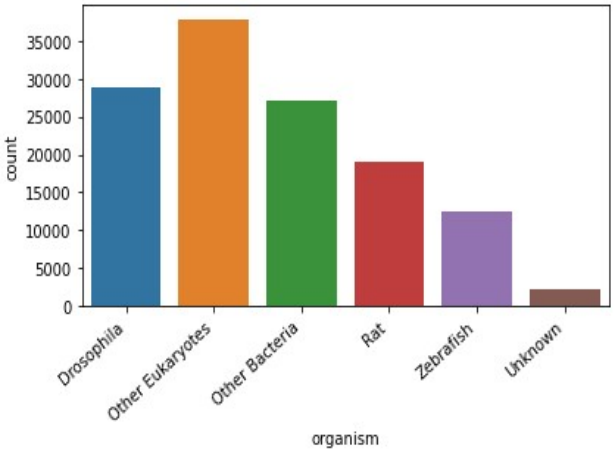
The CEM500k dataset consists of around 500,000 cellular level microscopic grayscale images taken from different organisms and captured with different kinds of microscope. Fig. 7 shows the distribution of the dataset in terms of types of organisms. In total there are 8 known types of organisms and a small portion contains organisms of unknown type. The diverse nature of the data set will allow the network to learn robust features.

Dataset	Train Size	Test Size	Modality	Usage
CEM500K	500,000	0	Microscopic	Pre-training
Lucchi++	165	165	Microscopic	Segmentation
Kasthuri++	85	75	Microscopic	Segmentation

TABLE I: Dataset Description: The table shows the different data sets and its properties. CEM500k contains 500000 images and it will be used for pre-training using BYOL Method. Other datasets will be used for segmentation task



(a)



(b)

Fig. 7: Distribution of CEM500K: CEM500k dataset consists of microscopy images from 10 different organisms. The organisms Mouse,Human and C. elegans constitute majority of the images. The dataset has been divided into two parts. Fig 7a shows the organisms comprising the first part of the dataset. This part has been used for the classification task. Fig 7b shows the organisms comprising the second part of the dataset. This part has been used in pre-training task.

Other than this, two benchmark datasets Kasthuri++ and Lucchi++ [2] have been chosen for the segmentation task. Both these datasets contain cellular level images of the mouse brain and the task is to segment mitochondria. Lucchi++ contains 165 training and testing images with annotations and Kasthuri++ contains 85 training and 75 testing images with annotations. Fig. 5a shows one of the images of Lucchi++ and fig. 5b shows the mask of the same image.

B. Implementation Details

- **Pre-training:**
  - **Data set:** As mentioned above CEM500k has been divided into two parts. Fig. 7b shows the distribution of the images used for BYOL pre-training task. Out of 500,000 which is the total size of the data-set, 150,000 have been used for the pre-training. As part of the pre-processing step, each gray-scale image have been resized to 128x128 px and concatenated across three channels to form a 3-channel image. Similarly, the colorized image is obtained by feeding the grayscale image to the colorizer network [29].
  - **Training:** Resnet50 and stand-alone self-attention [23] have been used as encoders in two separate pieces of training. The architecture of MLP is the same as that in the original BYOL method. The weights of the online network have been updated by back-propagating the loss calculated by comparing the output of online and target network. And the weights of the target network is the exponential average of the online network. Adam optimizer have been used with a constant learning rate of 0.0003. The training was first carried for 50 and then extended to 100 epochs. The batch size was set to 20. The parameters mentioned above were constant for resnet50 and stand-alone self-attention encoders.

- **Downstream Task:**
  - **Classification:** The subset of the CEM500k dataset shown in fig. 7a have been used to perform the classification task. Predicting which type of organism the image belongs to is part of this task. 100,000 images from this subset have been used for this task. The architecture of the classifier is shown in fig. 3a. Features extracted from the encoders (pre-trained using our method) are fed into the logistic regression model. Out of 100,000, 70,000 images have been selected for training, 10,000 for validatoin and 20,000 for testing. Cross-entropy loss function and adam optimizer with a learning rate of 0.0003 have been used to train the classifier. Accuracy have been used as a metrics to evaluate the classification results

and a confusion matrix have been used to visually understand the classification results of each encoder.

– **Semantic Segmentation:** Benchmark data sets Lucchi++ and Kasthuri++ have been used to perform semantic segmentation. Encoders derived from self-supervised pre-training have been incorporated into a U-net architecture. The dataset was divided by 75-25 split meaning 75% of the data was reserved for training and 25% for validation. The images were resized to 224 x 224 px and augmentations in the form of horizontal flip, vertical flip, rotation, gaussian blur were applied. The selection of the augmentations were random with the intention of robust training.
The results achieved were best when trained for 200 epochs with a batch size of 4. The loss function and optimizer used were dice-bce loss and adam optimizer with starting learning rate of 0.0005 and decreasing if there is no change in validation loss for 15 epochs.

C. Metrics

In this section we discuss about the metrics that were used to evaluate our results.

- mIoU: This metric is popular in semantic segmentation and it stands for mean Intersection over Union. First IoU is calculated for each class followed by the mean across all the classes. TP is the region that has been correctly predicted. FN is the region that belongs to the class but is incorrectly predicted as a different class and FP are those regions that belong to a different class but are predicted as the class.

$$IoU = \frac{TP}{TP+FP+FN}$$

D. Evaluation

The results obtained from our experiment were evaluated on the following parameters:

1) Compare the performance of the proposed method on benchmark data-sets: The U-net constructed from the encoders derived from our pre-training method have been tested on benchmark data sets Lucchi++ and Kasthuri++. Further, the results obtained have been compared against the results obtained in the paper [2]. The results obtained from this evaluation are shown in table V.

2) Compare the performance of encoders pre-trained using our method on the segmentation task: The purpose of this evaluation is to understand the difference in the performance of a segmentation model with and without our self-supervised pre-trained encoders. The encoders (resnet50 and stand-alone self-attention) derived from the pre-training have been used in U-net architecture to perform semantic segmentation. In this evaluation, the weights of both the encoder and decoder were updated through back-propagation. Similarly, U-net with resnet50 (Imagenet pre-trained) and traditional U-net have been constructed to perform segmentation and compare results with the byol pre-trained encoders. The results obtained from this evaluation is shown in table II.

3) Compare the representation learning capability of our self-supervised method against Original BYOL: The purpose of this evaluation was to quantify the representation learned using our method. This evaluation was carried out by testing the encoders on classification and semantic segmentation downstream tasks.

The encoders from our pre-trained method and the original BYOL were used to construct the Unet architecture for the task of semantic segmentation. The encoders were frozen(that means the weights of the encoder would not change during training) and the mIoU score was calculated to evaluate the results.

And for the task of classification, encoders from our pre-trained method, original BYOL, encoders from the Unet mentioned in the paper [2], resnet50 (Imagenet), stand alone self-attention (randomly initialized). Features have been extracted from each encoder and passed to the logistic regression model. The results from this evaluation is shown in table III and table IV.

VI. Results and Discussion

In this section we list out all the results for our experiment and discuss them in detail. The name of the encoders mentioned under U-net Encoder or Encoder column name in Table II, III, IV, V follow the nomenclature of Encoder name(ours), Encoder name(BYOL), Encoder name(traditional). Encoder name(ours) refers to the encoders used in self-supervised pre-training using our method. Encoder name(BYOL) refers to the encoder used in the self-supervised pre-training using the original BYOL method. An Encoder name(traditional) refers to the encoder either pre-trained on Imagenet or random initialised. Other than that we have also mentioned Encoder(Paper) which refers to the network mentioned in the paper [2]. This network is an improved version of original U-net in terms of parameters.

In our experiment we have used two encoders resnet50 which is pre-trained on Imagenet and stand alone self-attention [23] with randomly initialised weights.

A. Results

Table II shows the results obtained by encoders from our pre-training strategy and the obtained results have been compared against traditional U-net and resnet50 (Imagenet pre-trained) based U-net. The resnet-50 encoder pre-trained using our method achieves mIoU of 0.7896 for Lucchi++ and 0.77 for Kasthuri++. Stand Alone Self-attention-based pre-trained encoder on the other hand achieves 0.7325 and 0.75 on Lucchi++ and Kasthuri++ respectively.

On the other hand, traditional U-net achieves a mIoU score of 0.58 and 0.55 on Lucchi++ and Kasthuri++ respectively. And Resnet50 based U-net achieves mIoU score of 0.5821 and 0.62 on Lucchi++ and Kasthuri++ respectively. Overall our encoders performed better than the traditional U-net as well as the resnet50 based U-net. The difference in performance can be attributed to our pre-training method.

Dataset	U-net Encoder	mIoU
Lucchi++	Resnet50 (Our)	**0.7896**
	Self-Attention (Our)	0.7325
	U-net (traditional)	0.58
	Resnet50 (traditional)	0.5821
Kasthuri++	Resnet50 (Our)	**0.77**
	Self-Attention (Our)	0.75
	U-net (traditional)	0.55
	Resnet50 (traditional)	0.62

TABLE II: Comparison of Semantic Segmentation Results with and without pre-trained encoders: The table shows the results of the semantic segmentation on the Lucchi++ and kasthuri++. The encoders derived from pre-training are fine-tuned on both the datasets.

In Table III, the results obtained by our pre-training method is compared against the original BYOL pre-training method. The encoders from the pre-training were trained for semantic segmentation with frozen encoders. This means that only the weights of the decoder will be updated and that of the encoder will be same(frozen) throughout the training on segmentation datasets. Encoders Self-Attention and Resnet50 derived from our pre-training achieve mIoU score of 0.7034 and 0.6593 on Lucchi++ and 0.7167 and 0.6839 on Kasthuri++ data set.

Dataset	U-net Encoder	mIoU
Lucchi++	Resnet50 (Our)	0.6593
	Self-Attention (Our)	**0.7034**
	Resnet50 (BYOL)	0.6743
	Self-Attention (BYOL)	0.6530
Kasthuri++	Resnet50 (Our)	0.6839
	Self-Attention (Our)	**0.7167**
	Resnet50 (BYOL)	0.7036
	Self-Attention (BYOL)	0.6849

TABLE III: Comparison of our pre-training strategy with BYOL pre-training: The table shows the results of semantic segmentation when the encoders pre-trained with our pre-training strategy is compared against the BYOL. The weights of the encoders are not updated during the training on semantic segmentation dataset.

Further, the original BYOL [11] method was used to pre-train the two encoders (resnet50 and stand-alone self-attention) on the same data sets as our method. Resnet50 encoder achieves mIoU of 0.6743 and 0.7036 on Lucchi++ and Kasthuri++ respectively. Self-Attention encoder achieves mIoU of 0.6530 and 0.6849 on Lucchi++ and Kasthuri++ respectively.

The scores achieved by resnet50 encoder(original BYOL) based U-net is comparable on kasthuri++ dataset but lags behind the best performing attention model(pre-trained using our method) by 3% on Lucchi++. Stand-alone self-attention encoder from the original BYOL achieved a mIoU score of 0.6849 which is less than our self-attention results by 0.0318 or 3%.

The use of colorization in our pre-training method instead of a combination of augmentations gives comparable results on the Lucchi++ and is at times better than the original method(BYOL).

Next, we use the pre-trained encoders for the classification task and the results are shown in Table IV. The stand-alone self-attention model(our method) achieved an accuracy of 59.03%. Self-attention encoder derived from the original byol pre-training achieved an accuracy of 75.67% which is an improvement of 16.64% on our method. But the encoder from our method improves upon the performance of the randomly initialized stand-alone self-attention encoder by 3.71%.

Encoder	Accuracy(%)
Resnet50 (Our)	71.75
Resnet50 (BYOL)	72.3
Resnet50 (traditional)	70.715
Self-Attention (Our)	59.03
Self-Attention (BYOL)	**75.67**
Self-Attention (traditional)	55.32
Encoder (Paper)	36.83

TABLE IV: Classification Results on cem500k dataset: The table shows the results of classification on cem500k dataset.

The resnet50 encoder derived from our BYOL pre-training achieves an accuracy of 71.75% which is less than the performance of resnet50 derived from the original method by 0.8%. Again, our recent encoder performs better than the imagenet trained resnet50 by 1%. The encoder from the network proposed in [2] achieved an accuracy of 36.83% which is lower than our resnet model by 36.14% and the self-attention model by 22.2%. Fig 8 shows the way the resnet50 encoder(our method) shows the percentage of images classified accurately. It can be seen that the encoder is able to classify the mouse class with the highest accuracy and the C.elegans with the lowest accuracy. The miss-classification accuracy of the encoder does not go above 20%. The confusion matrix for self-attention encoders is shown in Fig 9. The encoder correctly classifies the human class with the highest accuracy and C.elegans class with the lowest accuracy.

In the end, the mIoU score obtained by the proposed method is compared against the results mentioned in the paper [2] as shown in Table V. Our resnet method achieves a mIoU of 0.7896 which is less than the results in the paper by 0.1564 and 0.15 for Lucchi++ and Kasthuri++. And the mIoU score of our self-attention model is less than the paper's [2] results by 0.226 and 0.17 on Lucchi++ and Kasthuri++ respectively.

Dataset	U-net Encoder	mIoU	mIoU(paper)
lucchi++	Resnet50 (Our)	0.7896	**0.946**
	Self-Attention (Our)	0.7325	
Kasthuri++	Resnet50 (Our)	0.77	**0.92**
	Self-Attention (Our)	0.75	

TABLE V: Comparison of segmentation results with benchmark results: The table shows the results obtained by our method and compared against the results obtained in the paper [2].

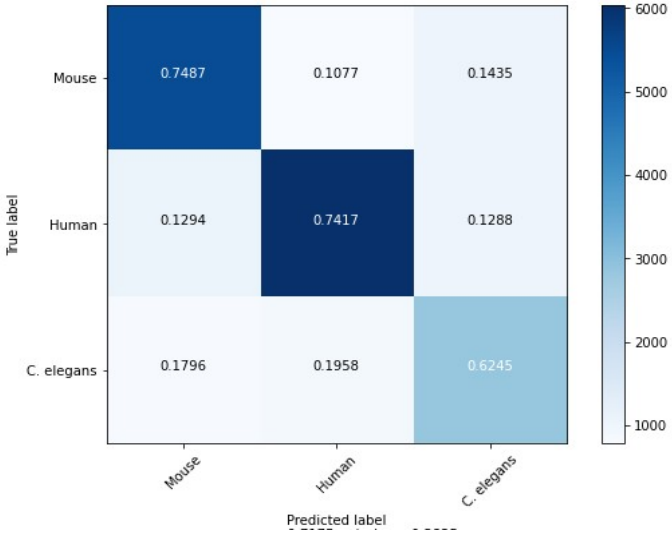


Fig. 8: Confusion matrix of our resnet50 pre-trained encoder: The figure shows the confusion matrix for classification by resnet50 derived from our byol method.

Overall, the resnet50 encoder works well in conjunction with our pre-training strategy across all the experiments. The self-attention model on the other hand works well on the semantic segmentation task but performs poorly on the classification task. The poor performance cannot be directly related to the architecture of stand-alone self-attention. This is because the same self-attention architecture has been pre-trained with the old byol method and has achieved the highest classification accuracy. Next, we analyze the CAM of the last layer of stand-alone self-attention pre-trained using the BYOL and our method. Fig 10 shows the microscopy image and Fig 11 shows the class activation map of the pre-trained encoders. The CAM of the self-attention encoder pre-trained using our method is shown in Fig. 11a. From the figure it can be seen that the encoder focuses on the cellular structures present in the image. Similarly, in Fig. 11b the activated region generated by the self-attention encoder pre-trained using BYOL is also some parts of a cellular structure. The regions activated by the self-attention model pre-trained using both the methods are visually similar.

A similar observation can be seen in the CAM of the resnet50 model. The CAM of the resnet50 pre-trained using our method as shown in Fig 11c activates a cellular structure in the image. And the resnet50 model pre-trained using BYOL

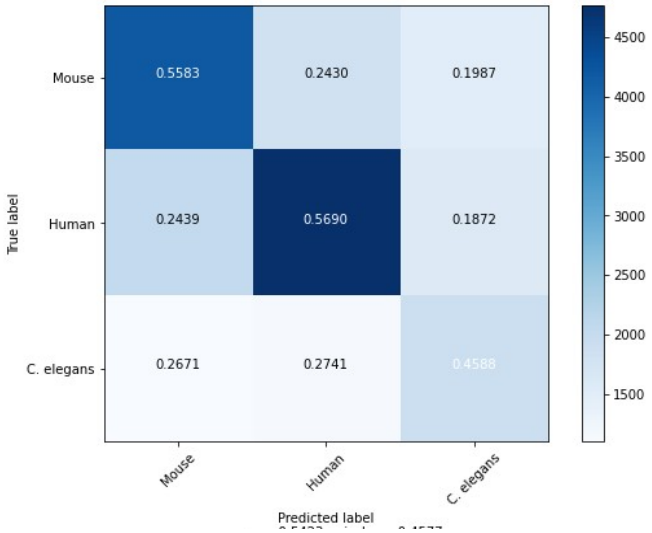


Fig. 9: Confusion matrix of our self-attention pre-trained encoder: The figure shows the confusion matrix for classification by stand alone self-attention derived from our byol method.

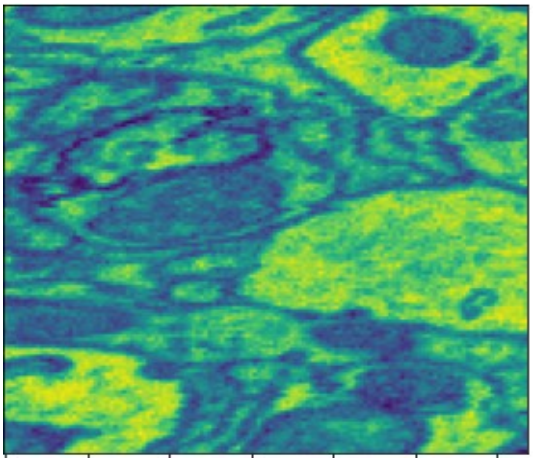
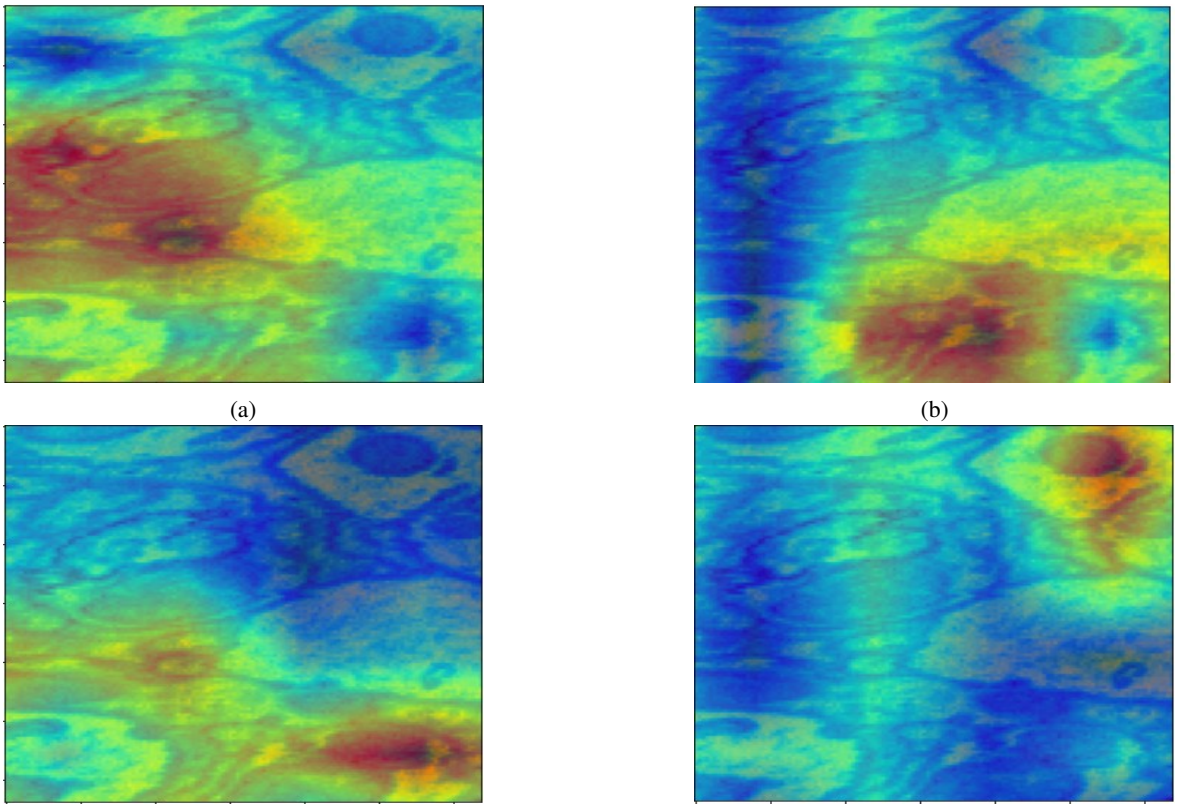


Fig. 10: Image: The figure shows the Image that is used to obtain the class activation map of the last layer of the pre-trained encoders.

also focuses on a cellular structure in a different part of the image. Overall, the investigation of CAM led us to believe that all the encoders activate visually relevant parts of the image.

VII. Conclusions

In this paper, we addressed the issue of scarcity of labeled images in the medical domain by proposing a self-supervised method for feature learning. Initially, we performed a literature review of the semantic segmentation methods. Further, we compared various self-supervised pre-text tasks and made a hypothesis that colorization tasks can be used to learn useful representation for microscopy medical images. Next, we reviewed contrastive learning methods and concluded that the BYOL method of contrastive learning is better than SimCLR in terms of efficiency and performance. We made another hy-



(a) (b) (c) (d)

Fig. 11: Class Activation Map: The figure shows the class activation map generated by the last layer of the pre-trained encoders. (a) In Fig 11a, the Class Activation Map of the self-attention model pre-trained using our method is shown. (b) In Fig 11b, the Class Activation Map of the self-attention model pre-trained using BYOL method is shown. (c) In Fig 11c, the Class Activation Map of the resnet50 model pre-trained using our method is shown. (d) In Fig 11d, the Class Activation Map of the resnet50 model pre-trained using BYOL method is shown.

pothesis about self- attention-based architecture that it would work better than convolution-based architecture owing to its ability to capture global features.

The encoders from our pre-training strategy were used to model an Unet. This Unet was further fine-tuned on two semantic segmentation benchmark data sets. Both the Unets (resnet50 and stand-alone self-attention) from our pre-training method performed better than the traditional U-net by a significant margin on both the data sets. Further, our encoders performed better than Imagenet pre-trained resnet50 based Unet. Overall, resnet50 (our method) based Unet performed better than the stand-alone self-attention (our method) based Unet by a score difference of 0.0571 on Lucchi++ and 0.02 on Kasthuri++. This result contradicts our assumption that self-attention based encoder will perform better than convolutional encoder.

Further, we evaluated the effect of colorization in our pre-training strategy by comparing it against the original BYOL method. The encoders were frozen and only the weights of the decoder part of the U-net were updated. Unet with stand-alone self-attention encoder pre-trained using our method performed best on both the data sets with mIoU score of 0.7034 and 0.7167 on Lucchi++ and Kasthuri++ respectively. The results of reset50 encoder pre-trained using our pre-training strategy and that from original BYOL were comparable with original BYOL getting better results by a very fine margin on both the datasets. These results help us conclude that the features learned by the self-attention encoder trained using our strategy is better than that of other encoders.

With this conclusion, we moved on to the next downstream task intending to make our conclusion full proof. In the classification task, stand-alone self-attention derived from the original BYOL performed best with the accuracy of 79.28% while our stand-alone self-attention scored 54.23%. On the other hand, resnet50 derived from our method scores 71.75% which is less than resnet50 derived from original BYOL by 0.595%. But at the same time, the results obtained by resnet50 are an improvement over the Imagenet pre-trained resnet50 by 1.04%. Also, both our models' resnet50 and stand-alone self-attention perform better than the encoder from the paper by 34.92% and 17.4% respectively.

Based on the two experiments, we get contrasting results. The self-attention gets good results on the semantic segmentation task but fails to classify the images with good

accuracy. So, we further analyze the CAM of the self-attention encoder to understand these contrasting results. We find that the features learned by all the encoders are visually similar and all the encoders activate important cellular structures in an image. Thus the dip in performance can be because of the hyperparameters of the logistic regression model.

Overall, the self-attention model pre-trained using our method performed better on semantic segmentation task when the weights of the encoder was not being updated(that is the encoder was frozen) as shown in Table III. And resnet50 pre-trained using our method, performed comparably to the best performing encoder on both the tasks. So, we conclude that inclusion of colorization task has assisted the encoders to learn relevant features and has performed better than the BYOL method on semantic segmentation task.

In the end, we compared our semantic segmentation results against the results obtained using the supervised learning approach. Our resnet50 based encoder achieved a mIoU score that was less than that of the supervised approach by 0.1564 on Lucchi++ and by 0.15 on Kasthuri++. The results obtained through our method are less in comparison to the supervised results. But, our encoders achieved better results on the image classification task than the encoder from the supervised approach.

With the results from our experiment, we answer the research question **Can a self-supervised learning method with a self-attention model develop a robust representation for semantic segmentation?** The stand-alone self-attention model pre-trained using our method was the best performing model on both the benchmark data-sets with frozen encoders and scored comparably with unfrozen encoders. In future work, different types of encoders like resnet200 can be pre-trained with different pre-text tasks instead of colorization to see if encoders learn better representation.

REFERENCES

[1] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1473–1481, 2020.

[2] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. *Nature Methods*, 16(12):1247–1253, Dec 2019.

[3] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[5] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[7] Andrei De Souza Inácio and Heitor Silvério Lopes. Epynet: Efficient pyramidal network for clothing segmentation. *IEEE Access*, 8:187882–187892, 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[13] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation. *CoRR*, abs/1811.11721, 2018.

[14] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *CoRR*, abs/1902.06162, 2019.

[15] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles, 2018.

[16] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. *CoRR*, abs/1603.06668, 2016.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[18] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2229–2235, 2018.

[19] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017.

[21] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

[22] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CoRR*, abs/1604.07379, 2016.

[23] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *CoRR*, abs/1906.05909, 2019.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[25] Sohil Shah, Pallabi Ghosh, Larry S Davis, and Tom Goldstein. Stacked u-nets: A no-frills approach to natural image segmentation. *arXiv e-prints*, pages arXiv–1804, 2018.

[26] Anil; Pang Guan; Torresani Lorenzo; Basu Saikat; Paluri Manohar; Jawahar C. V. Singh, Suriya; Batra. Self-supervised feature learning for semantic segmentation of overhead imagery. In *BMVC*, 2018.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[28] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification, 2017.

[29] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016.

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016.

[31] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.