Self-supervised Learning Through Colorization for Microscopy Images

Vaidehi Pandey, Christoph Brune, and Nicola Strisciuglio(✉)

Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands
n.strisciuglio@utwente.nl

Abstract. Training effective models for segmentation or classification of microscopy images is a hard task, complicated by the scarcity of adequately labeled data sets. In this context, self-supervised learning strategies can be deployed to learn suitable image representations from the available large quantity of unlabeled data, e.g. the 500k electron microscopy images that compose the CEM500k data sets.

In this work, we investigate a self-supervised strategy for representation learning based on a colorization pre-text task on microscopy images. We integrate the colorization task into the BYOL (Bootstrap your own latent) self-supervised contrastive pre-training strategy. We train the self-supervised architecture on the CEM500k data set of electron microscopy images. As backbone of the BYOL framework, we investigate the use of Resnet50 and a Stand-alone Self-Attention network, and subsequently test them as feature extractors for downstream classification and segmentation tasks.

The Self-Attention encoders pre-trained with the colorization-based BYOL method are able to learn effective features for segmentation of microscopy images, achieving higher results than those of encoders, both Resnet- and Self-Attention-based, trained with the original BYOL. This shows the effectiveness of colorization as pre-text for a downstream segmentation task on microscopy images. We release the code at https://github.com/nis-research/selfsup-byol-colorization.

Keywords: BYOL · Colorization · Microscopy images · Pre-training · Self-supervised learning

1 Introduction

Deep learning and convolutional networks achieved outstanding results in various computer vision and image processing tasks, such as image classification [11], object detection [1,24], semantic segmentation [3,25], place recognition [9,17], image and video generation [27,30], optical flow [12] and depth estimation [20], among others. In many cases, these models are trained using labeled samples in a supervised learning setting. For instance, semantic segmentation models [3,18,25,29] require images with pixel-wise labels: they are usually trained on data sets of natural images which contain large amounts of high quality accurately labelled

S. Sclaroff et al. (Eds.): ICIAP 2022, LNCS 13232, pp. 621–632, 2022.
https://doi.org/10.1007/978-3-031-06430-2_52

images. Examples are Cityscapes [7] or Mapillary Vistas [21], which contain images taken in cities. Collecting and labeling these images is time consuming, but it does not require particular expertise. In the case of medical or microscopy images, instead, acquiring a large number of images is prohibitive and labeling them requires expert knowledge. In [2], authors reported that experts spent 32 to 36 h annotating a microscopy data-set consisting of 165 images.

Recently, self-supervised learning demonstrated to be able to learn effective image representations from unlabeled data [4,5]. The self-supervision is created by defining an artificial pre-text task that exploits intrinsic structures in large amounts of unlabeled data, e.g. classification of image rotation/orientation, reconstruction from mosaique-images, etc. Encoder networks pre-trained with these techniques are then deployed as backbones for various computer vision tasks, and are either fine-tuned on a small amount of application-specific labeled data samples or directly used as feature extractors. In some cases, self-supervised pre-trained networks have achieved results comparable or higher than those of supervised networks [8,10].

In this work, we investigate using a colorization pre-text task in the BYOL pre-training framework to learn representations for microscopy cell image classification and segmentation in a self-supervised fashion. We exploit a large data set of unlabeled microscopy images, namely the CEM500k data set [6]. The use of colorization as pre-text task is motivated by the fact that it relates with shapes and regions of rather uniform color, that are also at the basis of image segmentation. It is thus expected that in the context of microscopy image analysis, this task can help learning some shape priors that could support further segmentation or classification tasks.

The rest of the paper is organized as follows. In Sect. 2, we provide a brief overview of related works while, in Sect. 3, we present our approach and model training strategy. In Sect. 4, we report the results that we achieved and finally draw conclusions in Sect. 5.

2 Related Works

Self-supervised learning methods leverage the data itself to disentangle data representation with no need of labels. The self-supervision is guaranteed by the design of pre-text tasks, which are artificial tasks to be solved by the network. In order to evaluate the quality of the learned representations, downstream tasks such as image classification or semantic segmentation are employed [13]. A good pre-text task is fundamental for self-supervised learning. The choice of the task determines the performance of the model on the downstream tasks. Some of the popular pre-text tasks are colorization [16,28], context prediction via image in-painting [23], jigsaw puzzle [15,22], image generation [31], among others.

The most powerful self-supervised learning methods are based on a pre-text task formulated as a contrastive learning problem, which consists of training two networks by forcing the representation of similar input image-pairs to be close in the latent space, and that of dissimilar input image-pairs to be distant in the

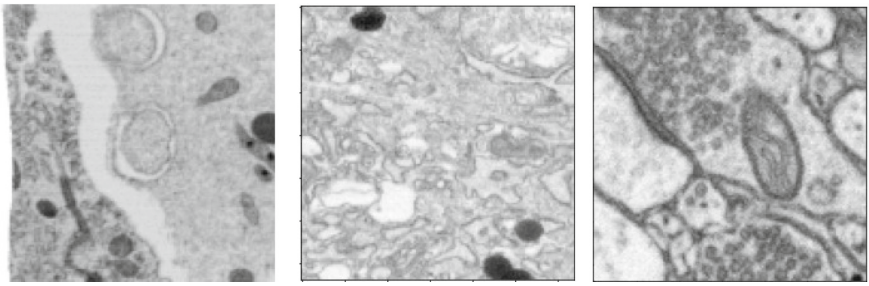
latent space. SimCLR [4] (Simple framework for Contrastive Learning) learns self-supervised visual representations by maximizing the loss between dissimilar images (negative pairs) and minimizing the loss between similar images (positive pairs). For each image in a training batch, two augmented versions are generated, which are considered as positive examples. The negative examples are the $2(N-1)$ images in the batch. In [8], it was observed that the performance of SimCLR is influenced by the choice of the augmentation pool for the pre-text task, and that removing the color distortion would result in a considerable drop of results. BYOL [8] discards dissimilar images pairs, making the training process more efficient. The representations learned in the contrastive architecture are processed through two different MLP networks, namely the online and the target network. The encoder in the online network branch is updated via stochastic gradient descent, while the decoder in the target network branch is updated using the exponential moving average (EMA) of the weights of the online network. A ResNet50 pre-trained with BYOL achieved 74% accuracy on ImageNet. Momentum Contrast (MoCo) [10] constructs a dynamic dictionary on-the-fly with a queue and average-moving encoder to support the learning of contrastive representations. It achieved competitive results on various computer vision tasks, namely image classification, detection and segmentation, substantially narrowing the gap with supervised methods. In [5], the authors explore a simplification of the siamese learning framework, called SimSiam, that does not rely on negative sample pairs, large training batches or momentum encoders. They propose a stop-gradient techniques to avoid collapsing solutions.

Self-supervised pre-trained models were deployed in semantic segmentation downstream tasks. Representations learned with BYOL were demonstrated to outperform other pre-trained ones by SimCLR and MoCo in semantic segmentation on the Cityscapes data set [7]. A modification of the in-paining pre-text task was proposed in [26], to overcome some of the limitations of the plain in-painting, which modifies the overall intensity of the input image by removing one or more patches. The use of an adversarial network to produce hard patches to in-paint demonstrated effective for pre-training of good representations for semantic segmentation, achieving higher performance than other methods on the Potsdam, SpaceNet and DG Roads datasets.

3 Data and Methods

3.1 Datasets

The CEM500k data set consists of about 500k electron-microscopy images containing structures at cellular-level, taken from different organisms and with different kinds of microscope. In Fig. 2, we show example images of cells from the organism classes c.elegans, human and mouse. We use images from these three classes to evaluate the performance of the pre-trained encoders on a downstream classification task. In Fig. 2, we show the distribution of the images in the data set, organized according to the type of organisms they are taken from. In total there are eight known types of organism, while a small portion of the data set



(a) c.elegans (b) human (c) mouse

Fig. 1. Example images from the CEM500k data set, taken from the classes of organism (a) c.elegans, (b) human and (c) mouse. We use a subset of these classes to test the downstream classification task.

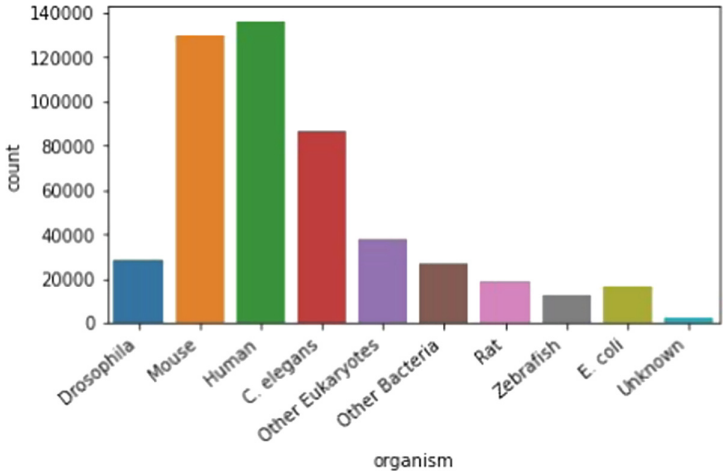
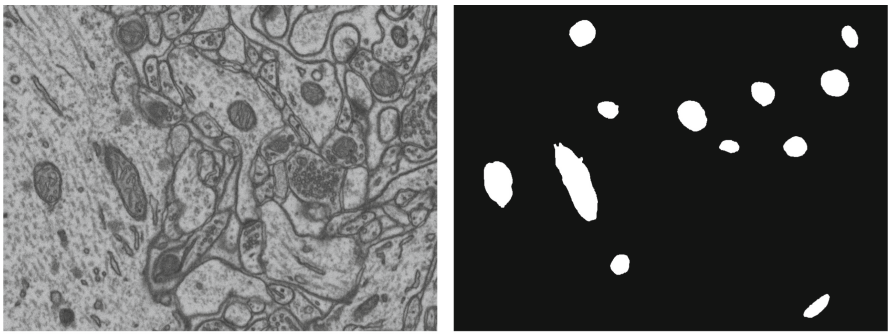


Fig. 2. Distribution of the images in different classes of organism in the CEM500k data set: images from eight known organism classes are present, plus a small portion of images for which the type of organism is unknown.

contains organisms of an unknown type. We use about 200k images for the self-supervised learning stage. The exploit the diversity of the images in the data set a to learn robust visual representations. For the downstream segmentation task, we use two benchmark data sets, namely the Kasthuri++ and Lucchi++ [2,14,19] data sets. They contain cellular-level images of the mouse brain, with labeled mitochondria regions. The Lucchi++ data set (a version of the EPFL Hippocampus dataset reannotated in [2]) contains 165 images with pixel-wise mitochondria annotations, while the Kasthuri++ data set contains 85 training and 75 testing images, also with mitochondria annotation. In Fig. 3a, we show one example image from the Lucchi++ data set, while in Fig. 3b we show the manually-made available ground truth mask of the same image.



(a) (b)

Fig. 3. An (a) image from the Lucchi++ data set, together with its (b) manual segmentation ground truth map.

3.2 Self-supervised Training

We use the BYOL framework [8] to train an encoder network for classification and semantic segmentation of microscopy images. We deploy the original BYOL architecture, see Fig. 4. It consists of 1) two encoders that share their weights, 2) an augmentation procedure and 3) a loss function. In BYOL we modify the generation of the augmented images for the target network, replacing the set of augmentation with an image colorization algorithm [28]. We make the code for experiments publicly available[1].

Encoder. The choice of the encoder is an important aspect of this method. In principle, one can choose any type of encoder architecture. In this work, we use a Resnet50 network and a Stand-alone Self-Attention network. While Resnet50 is a well-known convolutional network, the Stand-alone Self-Attention network is a custom modification of Resnet50. We substituted all the convolutional layers, except the first one, with self-attention layers. We thus investigate whether a different type of network, based on the self-attention principle, can be effectively used for self-supervised pre-training of methods for semantic segmentation and classification of structures in microscopy images.

Augmentation/Colorization. We use colorization as a pre-text task for the BYOL self-supervised learning framework. It converts a single-channel gray-scale image to a three-channel Lab image. We use the pre-trained colorization model proposed in [28] to convert the images in the CEM500k dataset from gray-scale to the Lab colorspace. A gray-scale and its corresponding color-augmented image form an input pair for the encoders as shown in Fig. 5. We call BYOL-colorization the method that we design using the colorization pre-text task, and BYOL-original the original version of BYOL, with an extended set of augmentations.

[1] Github repository: https://github.com/nis-research/selfsup-byol-colorization.

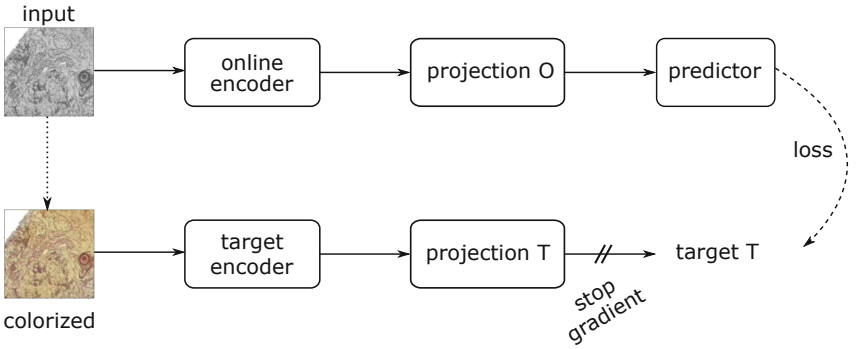


Fig. 4. Architecture of the BYOL framework that we used for self-supervised learning based on colorization pre-text. The online and target encoders have the same architecture but different weights. Only the online encoder and projection network O are trained via back-propagation (see stop gradient on the target network branch).

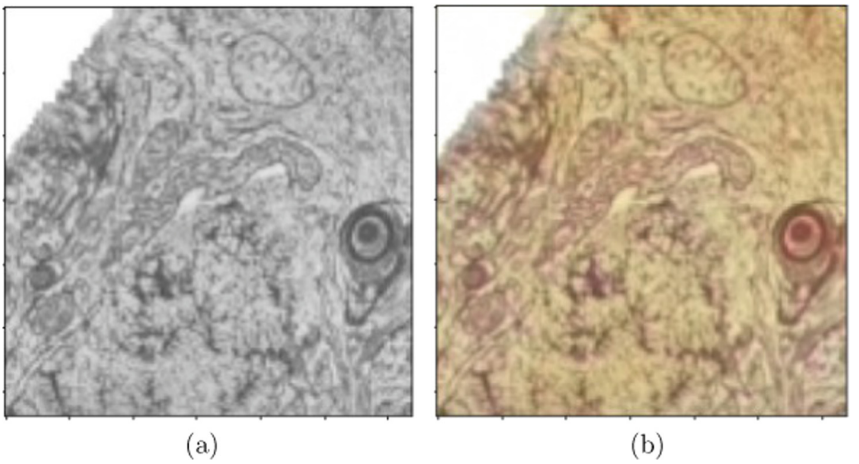


Fig. 5. An (a) example image from the CEM500k data set and (b) its colorized version obtained by using the model proposed in [28]

Loss Function and Training. The online and target encoder network in the BYOL architecture take as input the original gray-scale image and its colorized version, respectively. They have the same architecture but do not share weights. The target network provides the regression target to train the online network. The weights of the online network are optimized by back-propagating the gradient of an L2 regression loss function that compares the representations computed by the online and target network. The weights of the target network are updated as the exponential weighted average of the weights of the online network, according to the training scheme proposed in [8].

3.3 Downstream Tasks

We evaluate the image representation learned in the self-supervised pre-training on two downstream tasks, namely image classification and segmentation.

Classification. We use the pre-trained encoder as feature extractor in combination with a logistic regression model, to classify the images into three classes, namely *human cells, mouse cells, and c.elegans cells*. We train only the logistic regression model using a subset of the CEM500k dataset, which was not used for pre-training. In Fig. 6a, we depict the flow diagram of the classification downstream task.

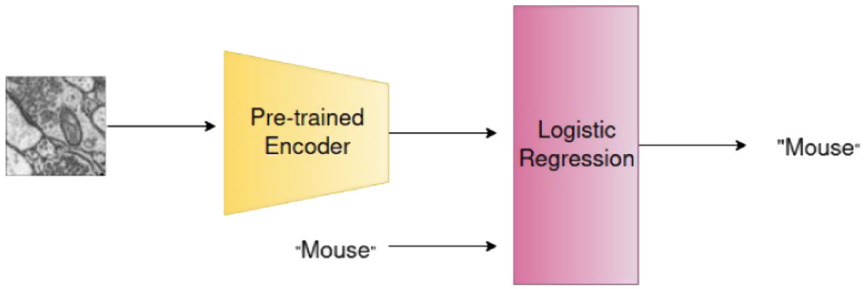
Semantic Segmentation. We considered semantic segmentation, namely the task of pixel-wise labeling image regions to belong to one out of a number of classes of interest, as another interesting downstream task to investigate for microscopy images. We use the pre-trained encoder as backbone for a UNet-like architecture, which contains a decoder network that computes a segmentation map of the same size of the input image. We compared the representation power of different backbones, pre-trained using BYOL-original, our BYOL-colorization, U-Net encoder and ResNet-50 pre-trained on ImageNet. We perform a fine-tuning stage, where the weights of the encoder stay unchanged while the weights of the decoder only are updated by back-propagation. In Fig. 6b, we show the flow diagram of the segmentation downstream task.

4 Experiments and Results

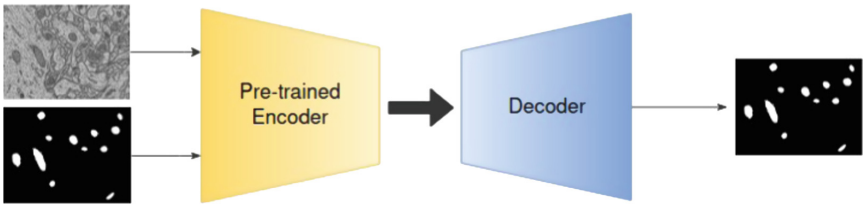
4.1 Experiments

We use the encoders pre-trained on the CEM500k data set as feature extractors for a classification and a semantic segmentation downstream task. For the classification task, we deploy our BYOL-colorization pre-trained encoders, namely the ResNet50 and Self-Attention networks, to extract features from images of a subset of the CEM500k data set. We then use these features together with a logistic regression classifier. We compare the performance of our encoders with that of similar encoders pre-trained with the original BYOL algorithm on the CEM500k dataset, and with a ResNet50 and a Self-Attention network pre-trained on ImageNet. Finally, we also use the encoder trained in [2]. While training the logistic regression classifier, we freeze the weights of the pre-trained encoders, so that we can test the effectiveness and quality of the pre-trained representations without adapting them to the downstream task.

Similarly, we compare the representation capabilities of our pre-trained encoders with those of BYOl-original pre-trained encoders on the task of semantic segmentation. Also for this experiment, we deploy the Resnet50 and the Self-Attention encoders as backbones. For evaluation purpose, we use the data sets proposed in [2,14] and [2,19], which contain segmentation labels. We freeze the encoder weights, and embed them into a U-Net architecture for segmentation, of which we fine-tune only the decoder part.



(a) Classification downstream task



(b) Segmentation downstream task

Fig. 6. Sketch diagram of the (a) classification and (b) segmentation downstream task. The weights of the pre-trained encoders (yellow boxes) are not updated while tuning the classifiers (red and blue boxes for classification and segmentation, respectiely) for the downstream tasks. (Color figure online)

4.2 Metrics

To assess and compare the performance of the pre-trained encoders on the classification downstream task, we computed the accuracy of classification. For the semantic segmentation task, instead, we measure the performance in terms of mean Intersection-over-Union (IoU) over the considered classes.

Table 1. Classification Results on cem500k dataset: The table shows the results of classification on cem500k dataset.

Encoder	Pre-training	Accuracy (%)
Resnet50	BYOL-colorization	71.75
Resnet50	BYOL-original	72.3
Resnet50	ImageNet	70.715
Self-Attention	BYOL-colorization	59.03
Self-Attention	BYOL-original	**75.67**
Self-Attention	ImageNet	55.32
[2]	Segmentation	36.83

4.3 Results

In Table 1, we report the results that we achieved on the downstream classification task. The Stand-alone Self-Attention encoder pre-trained with our BYOL-colorization achieved an accuracy of 59.03% while that pre-trained with the original BYOL achieved an accuracy of 75.67%. Our encoder improves upon the performance of the ImageNet Self-attention encoder by 3.71%. The Resnet50 encoder derived from our BYOL-colorization pre-training achieved an accuracy of 71.75%, which is slightly less than the performance of the Resnet50 derived from the BYOL-original pre-training by 0.8%. Our encoder performs better than the ImageNet pre-trained Resnet50 by 1%. The encoder of the network proposed in [2] achieved an accuracy of 36.83%, which is much lower than that of our BYOL-colorization pre-trained ResNet50 by 36.14%.

The BYOL-colorization pre-training allows to learn representations from unlabeled microscopy images which are more effective for classification than image representations learned on natural images from ImageNet. However, the diversity of data augmentation used in the original BYOL self-supervised pre-training approach allows to disentangle better features that are more effective for the classification task.

We report the results achieved on the downstream segmentation task in Table 2. We froze the weights of the pre-trained encoders and only trained the decoders for semantic segmentation. Self-Attention and Resnet50 encoders trained with our proposed BYOL-colorization pre-training achieved an mIoU score equal to 0.7034 and 0.6593 on the Lucchi++ data set, and equal to 0.7167 and 0.6839 on the Kasthuri++ data set. For both data sets, our BYOL-colorization pre-trained Self-Attention encoders achieved higher results than those of the BYOL-original pre-trained encoders. The results demonstrate that the use of colorization in our pre-training strategy contributes to learn suitable features for semantic segmentation of microscopy images. This is attributable

Table 2. 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.

Dataset	Encoder	Pre-training	mIoU
Lucchi++	Resnet50	BYOL-colorization	0.6593
	Self-Attention	BYOL-colorization	**0.7034**
	Resnet50	BYOL-original	0.6743
	Self-Attention	BYOL-original	0.6530
Kasthuri++	Resnet50	BYOL-colorization	0.6839
	Self-Attention	BYOL-colorization	**0.7167**
	Resnet50	BYOL-original	0.7036
	Self-Attention	BYOL-original	0.6849

to the fact that the colorization task induces the network to learn shape and color-region specific characteristics of the images, which better relate to the segmentation task. The wider range of augmentations learned in the original BYOL pre-training are not effectively tuned for segmentation.

The performance gap of pre-trained encoders for semantic segmentation of microscopy images with respect to supervised models is still large. The U-Net model adapted to the Lucchi++ and Kasthuri++ data sets proposed in [2] achieved an mIoU score (0.946 and 0.92) higher than that of BYOL-pre-trained encoder. In [2] the U-Net was trained for 1000 epochs, which is ten times larger than our 100 epochs fine-tuning of the decoder only, on the very few training images in the data sets, which may incur in overfitting, indicating that further investigation in the direction of evaluating the generalization properties of these networks is needed.

5 Conclusions

We investigated the feasibility of learning microscopy image representations from a large amount of unlabeled data in a self-supervised fashion. We thus address the problem of scarcity of unlabeled images, by training several Resnet50 and Self-Attention encoders using the BYOL self-supervised learning framework.

We demonstrated that using colorization as a pre-text task is effective to learn robust representations for semantic segmentation, and achieved better segmentation results than those obtained by encoders pre-trained using the set of augmentations designed for the original BYOL. For a classification downstream task, instead, the representation learned by the original BYOL showed slightly superior performance. The promising insights gained from the experiments open possibilities for further investigations in the direction of filling the performance gap between self-supervised and supervised methods for microscopy images, the latter of which may incur in overfitting caused by the long training schedules on very few labeled images.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. cite arxiv:2005.12872 (2020)
2. Casser, V., Kang, K., Pfister, H., Haehn, D.: Fast mitochondria detection for connectomics. Nat. Methods **16**(12), 1247–1253 (2019)
3. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR abs/1706.05587 (2017). http://arxiv.org/abs/1706.05587
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
5. Chen, X., He, K.: Exploring simple Siamese representation learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15745–15753 (2021). https://doi.org/10.1109/CVPR46437.2021.01549

6. Conrad, R., Narayan, K.: CEM500k, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. eLife **10**, e65894 (2021). https://doi.org/10.7554/eLife.65894
7. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
8. Grill, J.B., et al.: Bootstrap your own latent: A new approach to self-supervised learning (2020)
9. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-NetVLAD: multi-scale fusion of locally-global descriptors for place recognition. In: CVPR, pp. 14141–14152 (2021)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
12. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
13. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. CoRR abs/1902.06162 (2019). http://arxiv.org/abs/1902.06162
14. Kasthuri, N., et al.: Saturated reconstruction of a volume of neocortex. Cell **162**(3), 648–661 (2015). https://doi.org/10.1016/j.cell.2015.06.054
15. Kim, D., Cho, D., Yoo, D., Kweon, I.S.: Learning image representations by completing damaged jigsaw puzzles (2018)
16. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. CoRR abs/1603.06668 (2016). http://arxiv.org/abs/1603.06668
17. Leyva-Vallina, M., Strisciuglio, N., Petkov, N.: Generalized contrastive optimization of Siamese networks for place recognition. CoRR abs/2103.06638 (2021). https://arxiv.org/abs/2103.06638
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR abs/1411.4038 (2014). http://arxiv.org/abs/1411.4038
19. Lucchi, A., Smith, K., Achanta, R., Knott, G., Fua, P.: Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features. IEEE Trans. Med. Imag. **31**(2), 474–486 (2012). https://doi.org/10.1109/TMI.2011.2171705
20. Mayer, N., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE CVPR, pp. 4040–4048. arXiv:1512.02134 (2016)
21. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The Mapillary Vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017
22. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles (2017)
23. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. CoRR abs/1604.07379 (2016). http://arxiv.org/abs/1604.07379
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
26. Singh, S., et al.: Self-supervised feature learning for semantic segmentation of overhead imagery. In: BMVC (2018)
27. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: MoCoGAN: decomposing motion and content for video generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1526–1535 (2018)
28. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. CoRR abs/1603.08511 (2016). http://arxiv.org/abs/1603.08511
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. CoRR abs/1612.01105 (2016). http://arxiv.org/abs/1612.01105
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
31. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593 (2017). http://arxiv.org/abs/1703.10593