VPU Video Processing and Understanding Lab

UAM Universidad Autónoma de Madrid

SI1/PJI/2019-00414 AISEEME (2020-2022)

Aiding diagnosis by self-supervised deep learning from unlabeled medical imaging

D2 v1

D2 Enabling technologies: algorithms and findings

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

CONSEJERÍA DE EDUCACIÓN
E INVESTIGACIÓN

Comunidad de Madrid

AUTHORS LIST

Pablo Carballeira López	pablo.carballeira@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es

HISTORY

Version	Date	Editor	Description
1.0	10/01/2021	Pablo Carballeira López	First version
2.0	10/03/2021	Marcos Escudero Viñolo	Tasks descriptions and achievements.

CONTENTS:

1. Introduction

Convolutional neural networks (CNNs) have revolutionized artificial vision analysis as these networks yield close-to-human accuracy for challenging vision tasks by utilizing large and annotated datasets. However, generating and annotating these datasets are time consuming and sometimes require expensive expertise for some domains such as medical imaging. Self-Supervised Learning (SSL) has proven to be a successful strategy to tackle this problem. SSL does not use annotations; it generates pseudo labels by means of a pretext task (e.g., recognizing different augmented view of the same image) to train the CNNs high level semantics that are useful for solving vision tasks by re-training the CNN with small datasets.

This deliverable describes the work related to task T2.1: “Self-supervised frameworks and pretext tasks”, T2.2: “Skin lesion assessment” and T2.3: “Lung nodule malignancy evaluation”. The aim of T2.1 is to compare state-of-the-art SSL approaches, exploring the influence of the CNN architecture, the pretext task, and the training schedule. The aim of tasks T2.2 and T2.3 is to compare deep learning state-of-the-art approaches to skin lesion assessment and lung nodule malignancy evaluation.

2. Self-supervised frameworks and pretext tasks

2.1. Image embeddings for characterization.

We refer to an image embedding as the features extracted at a given layer of a deep learning model when a particular image is fed to it. These embeddings are accepted as a representative description of the image—subjected to the training target. Usually, one can expect that, at a given layer and for a given architecture, the higher the performance of the learned model is, the more representative the embeddings will be. A common way to obtain image embeddings is by using a network trained in the supervised mode [1][2].

Alternatively, SSL models can be used if images are to be represented in label scarce scenarios—as medical data requiring expert annotations or data acquired using devices capturing at non-visual modalities. SSL methods, instead of being trained for a label-driven task can be trained by using objectives such as a simple geometric task[3][4], pseudo labels generated through automatic clustering [5][6], or promoting proximity of "similar" data points in the feature space [8][9][10][11][12].

These objectives are commonly known as pretext tasks and can be used to arrange SSL models into three groups: geometric, clustering-based, and contrastive.

2.2. On the nature of pretext tasks.

Geometric models

One of the most straightforward approaches to defining a pretext task is applying a geometric transformation to an input image and training a network to solve it. The three geometric pretext tasks considered in this paper are rotation prediction [4], relative patch location prediction [3] and jigsaw puzzles [14]. The rotation prediction pretext task randomly applies one out of 4 rotations: 0º, 90º, 180º, 270º, to each training image sample and trains the network to predict which rotation was applied to a given image.

On the other hand, a model trained to predict patch locations is based on randomly sampling two close regions from an input image and training the network to predict their relative spatial location. Finally, when a jigsaw puzzle strategy is followed, the image is divided into tiles, that are then randomly shuffled. Then, the network is trained to predict their original arrangement.

Clustering-based representation learning

A more sophisticated approach to deep unsupervised learning is based on the classical clustering methods that are used to group unlabeled data into clusters according to some homogeneity criteria. An obvious way to incorporate clustering into the pretext task formulation is to perform clustering after each model update step. The generated labels are then used as pseudo-labels to evaluate the model in a supervised manner. These labels would, in turn, change the embeddings at the next step as the newly generated labels may differ from the labels at the previous step.

This is the strategy followed by Deep Clustering (DC) [5], that suffers from instability during the training process due to the random permutation of labels at each step. To tackle the issue of labels permutation and instability, Cluster Fit [15] relies on using a teacher network to define the pseudo-labels. Differently, in Online Deep Clustering (ODC) [6] the labels are updated using mini-batches and this process is integrated into the model update. This way, the embeddings and labels evolve together and the instability inherent in DC is eliminated.

Contrastive models

Top performing SSL models are driven by pretext tasks using contrastive losses [16]. Although exact implementations vary from model to model, the main idea remains the same: to learn representations that map the *positives* close together and push apart the *negatives*. The *positive* samples might be chosen based on modifications of patches in the same image or applying different augmentations obtained from the same image.

Non-Parametric Instance Discrimination (NPID) [17] treats each input image (instance) as belonging to a unique class and trains the classifier to separate between each instance via the noise-contrastive estimation [18]. The motivation for it comes from the observation that supervised learning approaches return similar embeddings for related images. Specifically, it is often the case that the second top scoring predicted class at the end of the model is semantically close to the first one following a human interpretation. Therefore, the network is expected to learn the semantic similarity between classes without explicitly having it as the objective.

Momentum Contrast (MoCo) [12] leverages a dynamic dictionary where *query* and associated *keys* represent image encodings obtained with an encoder network. If a *query* and a *key* come from the same image, they are a *positive* pair, otherwise a *negative* one. The *queries* and the *keys* are encoded by separate networks and the *key* encoder is updated as a moving average of the *query* encoder, enabling a large and consistent dictionary for learning visual representations.

Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [7], building on the principles of contrast learning, introduces a series of design changes that allow it to outperform MoCo [12] not requiring a memory bank. Among these changes are a more careful choice of data augmentation strategies, addition of a non-linearity between the embeddings and the contrastive loss, and increased batch sizes and the number of training steps. Further improving on the results of SimCLR [7], the second version of Momentum Contrast model (MoCo v2) [9] acknowledges its efficient design choices and takes advantage of an MLP projection head and more data augmentations.

Bootstrap Your Own Latent (BYOL) [11] reaches a new state-of-the-art on ImageNet linear classification while avoiding one of the greatest challenges that other contrastive models face: a need for negative pairs. BYOL circumvents this problem by generating the target representations with a randomly initialized model and then using them for its online training, by iteratively updating the target network, the online network is expected to learn better and better representations.

Finally, SwAV [19] describes a hybrid clustering-contrastive method that avoids the computation of pairwise distances between positive and negative samples by clustering the data in consistency-enforced clusters of the different image augmentations. Thereby, defining positive samples according to cluster memberships and reducing the distance storage requirements of the other contrastive methods.

2.3. Developing framework.

We have set up a very recent and powerful open-source framework for pretext-task comparison, the OpenSelfSup framework of the open-mmlab initiative [20]. This framework allows the definition of tailored pretext-tasks and self-supervised frameworks and already have trained models for some of the pretext tasks as well as for recent top-performing state-of-the-art ones, including deep clustering [6][19], instance discrimination [17], contrastive learning [7] [9] [12] and latent bootstrapping [19].

As a relevant outcome of the project, we have arranged this developing framework and include new architectures, datasets, and models for self-supervised learning. Trained models include those trained in the context of this project but also models from alternative existing developing frameworks [21]. Moreover, we have also defined protocols and created tutorials for including new ones in the future.

2.4. Study on the social-biased learned by self-supervised methods

Deep neural networks are efficient at learning the data distribution if it is sufficiently sampled. However, they can be strongly biased by non-relevant factors implicitly incorporated in the training data. These include operational biases, such as ineffective or uneven data sampling, but also ethical concerns, as the social biases are implicitly present—even inadvertently, in the training data or explicitly defined in unfair training schedules. In tasks having impact on human processes, the learning of social biases may produce discriminatory, unethical, and untrustworthy consequences. It is often assumed that social biases stem from supervised learning on labelled data, and thus, Self-Supervised Learning (SSL) wrongly appears as an efficient and bias-free solution, as it does not require labelled data. However, it was recently proven that a popular SSL method also incorporates biases. In this paper, we study the biases of a varied set of SSL visual models, trained using ImageNet data, using a method and dataset designed by psychological experts to measure social biases. We show that there is a correlation between the type of the SSL model and the number of biases that it incorporates (see a visual representation of it in Figure 1). Furthermore, the results also suggest that this number does not strictly depend on the model's accuracy and changes throughout the network. Finally, we conclude that a careful SSL model selection process can reduce the number of social biases in the deployed model, whilst keeping high performance.

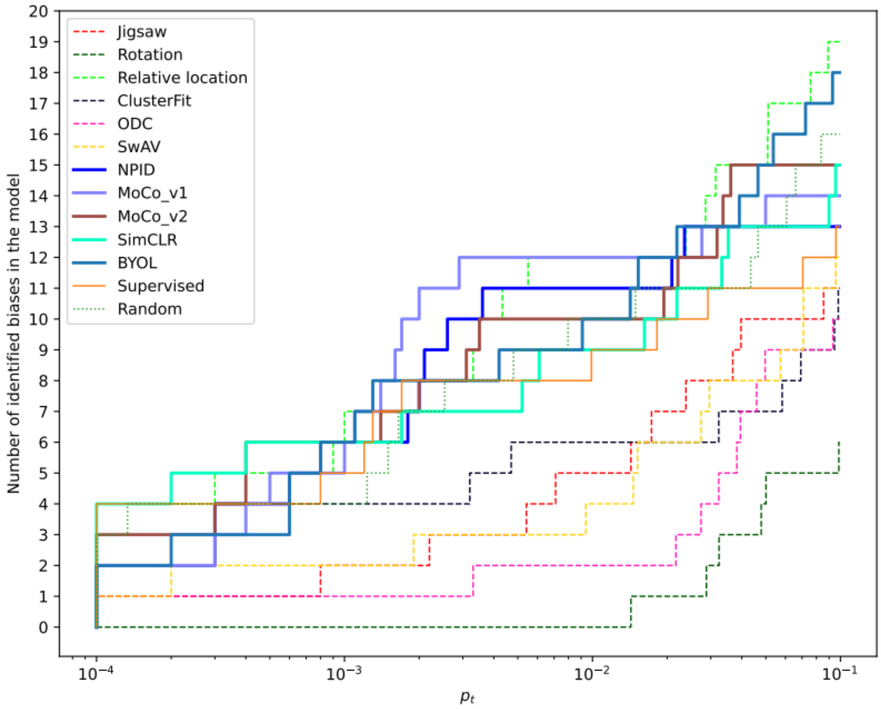

Figure 1. Number of biases detected in the embeddings of the Global Average Pooling layer for different values of pt. Biases detected for lower values of pt are statistically more significant. Contrastive models are plotted with thick solid lines, geometric mode

3. Skin lesion assessment

3.1. Initial findings

As a preliminary task, we have evaluated ([22][23][24]) SOTA methods on the ISIC 2017 and 2019 datasets [25]:

(1) Evaluated the effect of using a rotation prediction pretext task for learning features to be transferred for skin lesion assessment. Results suggest that rotation is a decent pretext task for learning representations on colour domains but fails in a domain less spatially structured as the skin lesion one [22].

(2) Evaluated the effect of using different CNNs architectures (ResNet, DenseNet and SqueezeNet) and data augmentation strategies. Preliminary results suggest that DenseNet yields top performance among the methods compared [23].

(3) Design a method to compare skin lesions in time, to quantify the evolution of the lesion and to provide additional data for the project. Preliminary results pave the road towards a semi-automatic analysis of these scenarios [24].

3.2. State of the art in skin lesion recognition

Due to privacy restrictions, to the relatively low rate at which certain medical conditions develop, and to the uneven distribution of suitable capture devices worldwide, a very small amount of medical imaging data is publicly available. In the subfield of skin lesion analysis, one of the largest accessible collections of imaging data is the International Skin Imaging Collaboration (ISIC) dataset that aggregates the data from other sources, such as HAM 10000 dataset [26], MSK dataset [27] and BCN 20000 dataset [28]. Annual public challenges based on the ISIC dataset target multi-class (predicting the exact type of a skin lesion) or binary (malignant vs. benign) classification problems. Whereas the latter can be considered an almost solved problem (ISIC-2020—a binary melanoma recognition problem—challenge winners achieved 0.949 on ROC AUC metric), the multi-class problem proposed in ISIC-2019 is still an open one, where the best reported approach [29] currently reaches only 72.5% ± 1.7% of balanced accuracy (average per-class accuracy).

Although ISIC-2019 is the largest publicly available skin lesion dataset with multi-class annotations, containing over 25000 labeled images, its size is small compared with those used for standard CNN training in well-established tasks. Moreover, the number of samples per class ranges from 239 to 12875, making it highly imbalanced and further complicating CNN training—e.g., a vanilla ResNet-50 [1] does not reach 50% of balanced accuracy, as shown in Table I. Some works address this issue by designing new loss functions that account for severe class imbalances [31]. However, the general trend seems to be increasing the complexity of neural models and utilizing deeper architectures, such as DenseNets [32] or very deep ResNets [1]. Continuing in the same direction, the top three best performing approaches in ISIC-2019 skin lesion diagnosis challenge are based on ensembles of neural networks that leverage multiple models to infer predictions [29], [33], [1].

3.3. Self-Supervised Learning in skin lesion recognition

The promise of learning useful representations without requiring labelled data led to the applications of SSL strategies to the medical imaging, where data labelling is challenging. Several recent works target the skin lesion recognition problem leveraging SSL approaches for model pretraining. This narrows the gap between SSL and fully supervised pretraining on ImageNet but, in most cases, does not yet close it. For example, a recently proposed approach [35] reaches 80.6% of accuracy on a multi-class ISIC-2018 classification problem by employing self-supervision to obtain transformation invariant features.

Specifically, features extracted at each epoch from the image decoder module of a CycleGAN architecture [36] are assigned to N clusters without a prior knowledge of N, using the maximum modularity clustering algorithm [37]. The memberships of the samples are used as pseudo labels to optimize the features.

A different study compares the individual performances of five existing SSL models [38] (BYOL [11], MoCo [12], SimCLR [7], InfoMin [39], SwAV [18]) on the ISIC-2019 data, but only considers the binary classification problem, reaching 0.956 on the standard for binary problems metric—ROC AUC (area under the precision-recall curve). On the contrary, another recent work does consider the multi-class skin lesion classification problem and shows that SimCLR pretraining outperforms supervised pretraining [40]. However, this study relies on a private dermatoscopic dataset containing over 450 000 samples.

Nevertheless, there is a gap in the current state-of-the-art as, to our knowledge, none of the existing approaches considers more than one SSL task in the pretraining stage, thus, limiting the resulting performance by not taking advantage of the pretext tasks of different nature. Moreover, with seldom exceptions, most works that use SSL pretraining in the skin lesion domain rely only on contrastive models (as these are generally the most accurate ones), thus, the potential contributions of clustering and geometric models are still barely explored. Finally, a multi-class skin lesion recognition problem is relatively unexplored by SSL methods, as most works focus on the binary melanoma recognition task, and only a few studies target the multi-class problem of the older (and smaller) versions of the ISIC dataset.

4. Lung nodule malignancy evaluation

The problems and uncertainty associated with the reference benchmark for this task [41] preclude the fulfilment of the project's objectives. Therefore, we moved to a similar non-colour modality task, in particular we focused on X-Ray chest [42] and COVID-19 [43] images.

4.1. On COVID-19 originated Pneumonias

The COVID-19 pandemic has had a devastating effect on the health and well-being of the world's population. One of the most important points in the fight against COVID-19 is the effective and early detection of infected patients, with medical imaging tests (X-rays and computed tomography) being one of the main forms of detection. In early studies [44] infected patients were found to have certain abnormalities in the chest (see Figure 2).

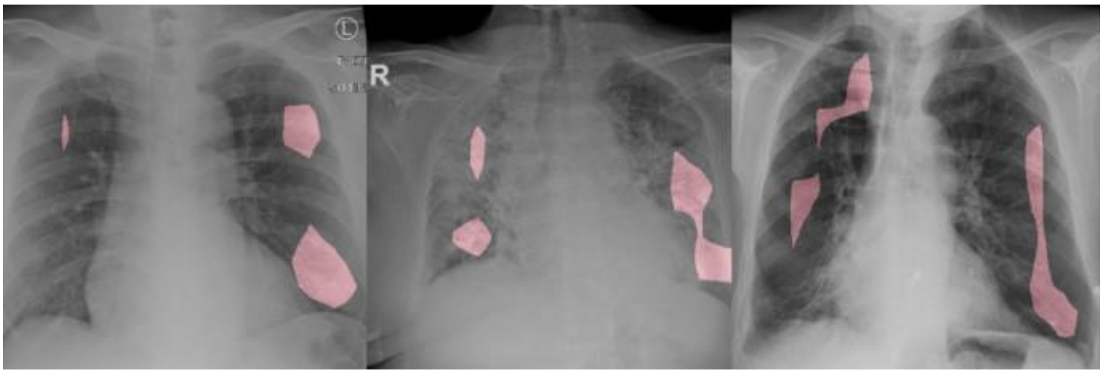

Figure 2 Example of abnormalities appreciable by x-ray images of different patients infected by COVID-19 and the associated critical factors (highlighted in red). Extracted from [45].

Although the diagnosis is mainly microbiological, imaging techniques play an important role in supporting this diagnosis, grading the severity of the disease, guiding treatment, detecting possible complications, and assessing the therapeutic response. Chest radiography is the first imaging method due to its wide availability and low cost. Thoracic computed tomography has greater sensitivity than chest radiography and allows assessment of both lung involvement and possible complications, in addition to providing alternative diagnoses. However, the latter is much more expensive and less accessible, so it has had less involvement in the studies.

4.2. Detection of COVID-19 originated Pneumonias in X-Ray images

Motivated by the need to achieve faster interpretation of medical images, several artificial intelligence systems based on deep learning have been proposed to accomplish this task. The conclusions of these have been quite promising in terms of detection of patients infected with COVID-19. Most of these studies have focused on the exploration of deep convolutional neural networks, thanks to the success that these algorithms present in artificial vision tasks.

Among these studies we can highlight COVIDNet [45], which will be studied in depth in Section 3.3. It is a deep convolutional neural network, which uses chest X-ray images to carry out detection of possible cases of COVID-19. This model allows the images to be classified into three classes, separating the images according to whether they present pneumonia caused by the COVID-19 virus, whether it is pneumonia caused by any other virus, or the patient in question does not present any type of pneumonia.

Another of the main studies carried out for the detection of COVID-19 through medical imaging is Deep COVID [46]. This study is based on the use of a deep learning framework that directly predicts possible COVID-19 infections from raw images, without the need to perform prior feature extraction. It is made up of four well-established and studied convolutional neural networks

(ResNet18 [1], ResNet50 [1], SqueezeNet [47], and DenseNet161 [32]). The results obtained for the four neural networks are included in Table 1.

It is also worth noting the study carried out by a group of Brazilian researchers [48], in which they intend to obtain an accurate and efficient method in terms of memory and processing time for the detection of COVID-19 in chest X-rays. To do this, they use the family of convolutional neural networks EffcientNet [49], known for the high precision it achieves with great efficiency.

ConvNet	Accuracy(%)
ResNet18	86.9
ResNet50	89.9
SqueezeNet	89.7
DenseNet161	86.3

Tabla 1. Results obtained by the Deep-COVID model for different types of ConvNet architectures. Adapted from [46].

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, 2016.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[3] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision, pages 1422–1430, 2015.

[4] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations (ICLR), 2018.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 132–149, 2018.

[6] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised epresentation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6688–6697, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 22243–22255. Curran Associates, Inc., 2020.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training -supervised visual transformers. arXiv preprint arXiv:2104.02057, 2021.

[11] Jean-Bastien Grill, Florian Strub, Florent Altch´e, Corentin Tallec, Pierre Richemond,

Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

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

[13] In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[14] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In European conference on computer vision, 2016.

[15] Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6509–6518, 2020.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006.

[17] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3733– 3742, 2018.

[18] Michael Gutmann and Aapo Hyv¨arinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[19] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems, volume 33, 2020.

[20] Xiaohang Zhan, Jiahao Xie and Enze Xie, OpenSelfSup, GitHub - open-mmlab/OpenSelfSup: Self-Supervised Learning Toolbox and Benchmark. 2021.

[21] Priya Goyal and Quentin Duval and Jeremy Reizenstein and Matthew Leavitt and Min Xu and Benjamin Lefaudeux and Mannat Singh and Vinicius Reis and Mathilde Caron and Piotr Bojanowski and Armand Joulin and Ishan Misra. VISSL: https://github.com/facebookresearch/vissl. 2021.

[22] Alejandro Camacho Valladares, "Aprendizaje auto supervisado para reconocimiento de objetos". Trabajo Fin de Grado (Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid), Junio 2020.

[23] Francisco Javier Martín Ameneiro, "Detección precoz de cáncer de piel en imágenes basado en redes convolucionales", Trabajos Fin de Grado (Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid), Junio 2020.

[24] Juan Antonio Álvarez Castillo, Análisis de la evolución, en número y tamaño, de lesiones de piel en zonas amplias del cuerpo, Trabajo Fin de Grado (Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid), Junio 2020.

[25] Codella, Noel CF, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the

international skin imaging collaboration (isic)."In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168-172. IEEE, 2018.

[26] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018; sci data (5): 180161,” 2018.

[27] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler et al., “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018, pp. 168–172.

[28] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig et al., “Bcn20000: Dermoscopic lesions in the wild,” arXiv preprint arXiv:1908.02288, 2019.

[29] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution efficientnets with meta data,” MethodsX, vol. 7, p. 100864, 2020.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] T. D. Tô, D. T. Lan, T. T. H. Nguyen, T. T. N. Nguyen, H.-P. Nguyen, L. Phuong, and T. Z. Nguyen, “Ensembled skin cancer classification (ISIC-2019 challenge submission),” ISIC Challenge, 2019.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[33] S. Zhou, Y. Zhuang, and R. Meng, “Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles,” ISIC Challenge, 2019.

[34] F. Pollastri, J. Maro˜nas et al., “Aimagelab-prhlt at isic challenge 2019,” AImageLab, Tech. Rep, 2019.

[35] D. Wang, N. Pang, Y. Wang, and H. Zhao, “Unlabeled skin lesion classification by self-supervised topology clustering network,” Biomedical Signal Processing and Control, vol. 66, p. 102428, 2021.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[37] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On modularity clustering,” IEEE transactions on knowledge and data engineering, vol. 20, no. 2, pp. 172–188, 2007.

[38] L. Chaves, A. Bissoto, E. Valle, and S. Avila, “An evaluation of self-supervised pre-training for skin-lesion analysis,” arXiv preprint arXiv:2106.09229, 2021.

[39] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” in Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., 2020, pp. 6827–6839.

[40] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen et al., “Big selfsupervised models advance medical image classification,” arXiv preprint arXiv:2101.05224, 2021

[41] Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao et al. "The lung image database

consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." Medical physics 38, no. 2 (2011): 915-931.

[42] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097-2106. 2017.

[43] Zhao, Jinyu, Yichen Zhang, Xuehai He, and Pengtao Xie. "Covid-ct-dataset: a ct scan dataset about covid-19." arXiv preprint arXiv:2003.13865 (2020).

[44] Chamorro, E. Martínez, A. Díez Tascón, L. Ibáñez Sanz, S. Ossaba Vélez, and S. Borruel Nacenta. "Radiologic diagnosis of patients with COVID-19." RadiologíA (English Edition) 63, no. 1: 56-73. 2021.

[45] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images." Scientific Reports 10.1: 1-12. 2020.

[46] Minaee, Shervin, et al. "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning." Medical image analysis 65: 101794. 2020.

[47] Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." *arXiv preprint arXiv:1602.07360*. 2016.

[48] Hu, Mu, et al. "Penet: Towards precise and efficient image guided depth completion." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

[49] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.