Skin Cancer Diagnosis Using Self-Supervised Learning

Maria Rita Ribeiro da Fonseca Verdelho
Instituto Superior Técnico - Universidade de Lisboa, Portugal
Email: ritaverdelho@tecnico.ulisboa.pt

***Abstract*—Convolutional Neural networks (CNNs) are the standard approach for image classification. However, they require a large amount of data and corresponding annotations. Collecting medical data is a difficult task, due to privacy restrictions. Moreover, it is even harder to obtain the clinical labels, since these must be provided by specialists. Self-supervised learning (SSL) has emerged as a possibility to overcome this issue, since it uses non-annotated data to pre-train the CNN. Recently SSL has been applied in the context of skin cancer. However, the results were not conclusive since a qualitative analysis was missing. Moreover, a proper analysis of the impact of different SSL approaches is still missing. In this master's thesis it will be investigated two SSL approaches: Rotation and SimCLR. The results highlight the benefits of applying self-supervised learning to the classification of dermoscopy images. Additionally, it was possible to demonstrate that these approaches learn different and complementary features, which is also a novelty of this thesis. As SSL is known to benefit from using more unlabeled data, it was also studied the impact of adding more data to the SSL pre-trained models (using 50% more data). It was possible to observe that depending on the level of difficulty of the task, the more it benefits from using more data. Therefore, the SimCLR task benefited more from the increase of data. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.**

***Index Terms*—Skin Cancer, Deep Learning, Self-Supervised Learning, Dermoscopy.**

I. Introduction

Skin cancer is one of the most common types of cancer worldwide [1]. In the past decade, the number of melanoma cases diagnosed has increased by 47% and in non-melanoma cancer about 5.400 people worldwide die every month due to this disease. Skin cancer is also one of the most treatable forms of cancer when detected in an early stage. However, late detection can have a significant impact on mortality rates. Therefore, there is a need to develop a convenient and precise method to perform early diagnosis [2].

Over the past decade, convolutional neural networks (CNNs) have been developed to assist human experts and accelerate the process of skin cancer diagnosis [2]. However, these methods require a huge amount of annotated data to obtain satisfactory results. Collecting medical data is a difficult task, due to privacy and law restrictions, and it is even harder to obtain clinical annotations since these must be provided by specialists [3]. To overcome this issue, the research community has been relying on transfer learning (TL). This method consists of first training a model for a task using a large data base and then "recycle" it for a new target task [4]. These pre-trained models usually have deeper architectures than what is needed in medical image analysis [5]. Additionally, the color distribution of natural images is also very different from the medical ones [6], which can result in models that have difficulties in generalizing to the other data [5].

Self-supervised learning (SSL) has emerged as a strategy to avoid the annotation process. This technique takes advantage of unlabeled data to perform a pre-training of the CNN [7] [8], allowing the model to learn relevant image features that can later be applied to a specific task. Recently, SSL has been used in the skin image context. Li *et al.* [8], Tajbakhsh *et al.* [5], Kwasigroch *et al.* [3] and Chaves *et al.* [9] applied different SSL techniques to the skin cancer diagnosis. Despite the promising results, it is still unclear which is the best SSL strategy for skin images. Additionally, all works focus solely on a quantitative analysis, disregarding the impact of SSL on the features learned by the model. However, in addition to the quantitative analysis that all works tend to have, this thesis introduces a qualitative assessment of the impact of the different pre-training strategies. Two different SSL techniques are also combined during this work: geometric distortion and contrastive learning.

This thesis aims to shed a new light on the application of SSL in the skin cancer context. Therefore it was developed a framework to: i) investigate the impact of SSL on the training and generalization of a CNN for skin lesion diagnosis and demonstrate that even with a small dataset there are benefits in using SSL; ii) compare two different SSL approaches; iii) for the first time provide a qualitative assessment of the impact of the different pre-training strategies, using explainability approaches; iv) demonstrate the complementarity of the features learned by the SSL strategies and the benefits of combining them; v) investigate the impact of adding more unlabeled data to the SSL techniques.

This is believed to be the first work to perform a robust quantitative and qualitative validation of the impact of SSL, and to demonstrate the importance of combining different SSL techniques.

The remaining of the paper is organized as follows. Chapter II contains a brief explanation of the background of deep learning; III introduces the used methodologies, and Chapter IV describes the experimental setup and the results and Chapter V concludes the paper.

II. Background

This chapter contains a brief explanation of the background of deep learning. Starts with a general description of CNNs. It is followed by an explanation of the supervised learning technique, which is currently the most common technique used in skin cancer diagnoses; explains SSL and addresses the differences between SSL and TL.

A. Convolutional Neural Networks - CNNs

CNNs have many applications from image recognition to image classification or object detection among others [7]. These networks receive an image as input and assign different importance values, given by learnable weights and biases, to multiple objects in the image. These parameters allow the network to distinguish different images [10].

1) ***Basic Concepts****:* CNNs receive images as input and submit them to a series of convolutional layers with filters, also known as kernels, followed by a non-linear activation function, in order to extract features from the images. The output of each layer is known as the feature map, which consists of an image different from the original. The feature maps will be submitted to a pooling layer that allows the CNN to reduce the dimension of each image. This process is repeated as many times as needed. Finally, a fully connected layer (FCL) is applied to convert the feature maps into a single array [10]. The set of FCLs is known as softmax classifier, which performs the intended multi-class classification by assigning a probability of each class label over all the classes [11].

2) ***Training the model****:* The training phase of the CNN has the aim of optimizing the model's weights, to allow the network to better map the input to the correct predicted class [10]. A loss function is used to improve the quality of these output predictions by comparing the predicted output to the true label. Many different loss functions have different objectives. The training phase can be seen as an optimization problem, where the minimum of the loss function is being searched. The network parameters are optimized through the gradient descent method, which indicates the right direction for the next iteration, in order to achieve the minimum of the loss function.

There are two phases when training the network. The forward phase, where the input goes through the network and the backward phase, where the gradients are back propagated and the weights are updated [10]. The latter phase is where the gradient of the loss function is calculated. The weights initialization is a hyper-parameter of the network. The choice of this initialization is typically done, in supervised learning, by training the network from scratch or by using transfer learning with pre-trained models [11]. By computing the forward phase an output is obtained and a loss is computed. The back propagation phase initiates and the gradient of the obtained loss function is computed. To reduce the loss function value the weights are updated.

3) ***ResNet Architecture****:* All the experiments carried out in this thesis use the ResNet-50 architecture. The work presented in [12] introduces the concept of a residual neural network that aims to facilitate the training of convolutional neural networks. In the past, it was proven that with the increase in the depth, the accuracy of the model tends to saturate and, then, degrades rapidly. In other words, by adding more layers into a previouly trained network there is a decrease in the accuracy of the model. To avoid this problem, instead of staking layers directly, this paper proposes a novel solution that consists of replacing the traditional convolution blocks with residual connections. These residual connections can be seen as 'shortcuts' that can be directly used once the input and the output have the same length. As ResNet has proved to be a less complex network and nevertheless it still manages to obtain good results this is why it will be used during this thesis work.

B. *Supervised Learning*

Over the last three decades, there has been an effort towards the development of machine learning methods to detect and classify the different skin cancer lesions. These methods are being created in order to help dermatologists correctly diagnose the different lesions [8]. The question that now arises is: What are the most popular methods to diagnose skin cancer lesions?

In the context of skin cancer, supervised learning is the most common approach. In order to classify different skin lesions, the images from the dataset are considered as features and the medical annotations associated to each image are the labels. The network is trained using labeled images and it is expected to acquire knowledge from the dataset in order to generalize the learned information to the new input images [13]. This confirms that the main supervised learning problem resides in the collection of a large dataset [14]. However, these large amounts of data are not easy to collect. Obtaining medically labeled images is even harder [2].

In supervised learning, the choice of the weights initialization is typically done by training the network from scratch or by using TL with pre-trained models. Training from scratch resides in assigning arbitrary weight values to the system, which means a new model is being constructed. On the other hand, TL uses weights that have already some image knowledge [4]. Therefore, this technique avoids the use of huge data, which resulted in an easier and faster method when compared to training the network from scratch [11].

1) ***Transfer Learning (TL)****:* TL, as the name indicates, uses the foundation of exporting knowledge from one task to another. This technique uses a model already pre-trained in a labeled dataset and 'recycles' some of the initial convolutional layers that have acquired some knowledge and train the rest of the layers to adjust to the new target task. This way, the network begins with weights that have already some image knowledge that has little similarities to the medical images. TL can be very advantageous. However, once this technique requires the use of a non-related dataset the learned weights can have problems generalizing well enough to the target tasks and datasets. Since the classes from both tasks are very different [15]. Another visible limitation is the fact that there is still the need to use labels in the pre-training phase.

Thus, other questions arises: what if we could join the supervised learning technique with the self-supervised, which

is currently gaining popularity, and apply it to skin cancer diagnoses? This could solve the TL limitations.

C. Self-Supervised Learning (SSL)

SSL was created to optimize the data usage since this technique does not require the use of labels in the pre-training phase [8]. Therefore, it extracts visual features from the unlabeled data [3]. The main goal is to use the learned weights to initialize a CNN for a specific target task, which is, in the skin cancer image analysis, the classification of the different skin lesions. To achieve this goal, the model is trained to execute a simple task, known as pretext task, for which labels can be easily generated without human supervision. Pretext tasks aim to extract different feature representations from the images. Therefore, it is important to select a SSL technique (between Geometric Distortion, Patch Relative Position, Colorization, Generative Modeling and Contrastive Learning) that is adequate to the wanted target supervised task. SSL can be divided into two steps: i) **Pretext task:** where the network has the ability to learn a new task using unlabeled data; ii) **Supervised Target Task:** consists of training a labeled dataset (with fewer annotated data) on the target task (which is in this case image classification) using the knowledge obtained by the pretext task [6].

*1) **TL vs SSL**:* SSL is similar to TL, but instead of pre-training a network using a labeled dataset, it uses an unlabeled dataset and extracts feature representations from the images by forcing the network to execute simple tasks. While executing these simple tasks the network learns parameters that are fine-tuned on the target task. In other words, the weights obtained during the visual feature extraction phase are then used to initialize the convolutional layers of the CNN. Therefore, SSL recycles the first convolutional layers of the pre-trained network (trained on the unlabeled dataset) and adjusts the rest of the layers to the new target task.

The main question that now arises is: Does it make sense to apply SSL to the skin cancer diagnosis? This question is addressed below.

*2) **Skin Cancer Diagnostic**:* Recently, SSL has been used in the skin image context. However, it is important to stress that most SSL techniques are very recent and, consequently, there are still few works that use them. Both Li *et al.* [8] and Tajbakhsh *et al.* [5] applied SSL techniques with color-based pretext tasks to the segmentation of skin lesions. Kwasigroch *et al.* [3] applied two SSL techniques based on geometric distortion to the skin cancer classification task. The closest work to the one executed in this thesis is that of Chaves *et al.* [9], in which they assess five SSL contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and improving model accuracy. Despite the promising results, it is still unclear which is the best SSL strategy for skin images. Additionally, all works focus solely on a quantitative analysis, disregarding the impact of SSL on the features learned by the model.

III. Methodologies

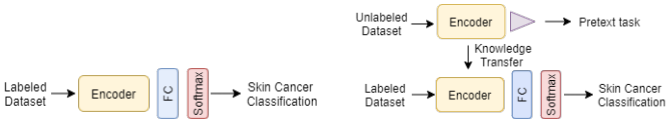
This chapter gives a brief explanation of the two strategies adopted in this work, as well as the experimental setup adopted in the skin cancer problem.

A. Proposed Approach

This thesis aims to perform a robust assessment of the impact of SSL as a pre-training technique, to initialize the weights of a CNN for skin cancer diagnosis. To better understand the impact of SSL, there was performed a systematic assessment, adopting the following pipeline:

(i) **Baselines** - two standard supervised learning strategies, where the weights of the CNN are initialized either at random (trained from scratch) or using a pre-trained model on ImageNet (fine-tuning).
(ii) **Scratch + SSL** - standard SSL methodology, where the weights of the CNN are initialized at random and refined using either the Rotation or the SimCLR technique.
(iii) **ImageNet + SSL** - a variant of the SSL approach, that aims to leverage the information from a model pre-trained on the ImageNet dataset. Here, the weights of the model used in the SSL phase are initialized from ImageNet and, then, they are refined using either the Rotation or SimCLR approach.
(iv) **Fusion** - fusion of the CNNs pre-trained using the Rotation and SimCLR techniques both at the feature (early fusion) and classification (late fusion) level.

Fig. 1 (a) describes the proposed generic approach for the application of supervised learning (baselines) and Fig. 1 (b) describes the proposed approach for the application of SSL. For the latter, the first step consists of pre-training the CNN using the chosen pretext task and, secondly, fine-tuning the parameters of the model to the classification task (this time using labels), by recycling the encoder and adding a FCL to output the 8 classes presented in the used skin cancer dataset. In all experiments, the encoder is a ResNet-50 [12].



(a) Supervised Learning. (b) Self-supervised Learning.

Fig. 1. Proposed framework using different initialization techniques applied to the skin cancer diagnoses. In both models, the last layer is a fully-connected one with 8 units. The triangle represents the last layers of the CNN specific of the pretext-task.

B. Data and Training Manipulation

During the execution of this thesis some issues needed to be corrected both in data and training.

*1) **Image Pre-Processing**:* The images presented in the ISIC archive were collected at different medical centers (each center generated images with different sizes, colors, and aspect ratios). Therefore, it was necessary to pre-process them. This process compensated the color and allowed all the images to have the same size while maintaining their aspect ratio. After having resized all the images to the desired size (224x224), it was applied the color constancy algorithm Shades of Gray as it is proposed in [16].

*2) **Training Specifications**:* In order to improve both the performance of the supervised and self-supervised classifiers, there has been used the technique of artificially augment the training set which prevents overfitting. This technique creates more variability in the data. To do so random flips (both horizontal and vertical) and rotations of multiples of 90 degrees were performed to all the images presented in the training set. These geometric transformations result in an augmentation of the training dataset, which allowed the network to have better performance.

The used dataset is highly imbalanced, in order to overcome this issue there have been applied class weights to the loss function. This technique assigns to the less frequent classes the higher weight and therefore the loss becomes a weighted average. This allows the model to be more robust since it does not tend to classify all classes with the category that appears more frequently in the dataset. Therefore it promotes a classifier that can learn all classes equally.

C. Initialization Techniques

This thesis aims to verify the impact on the application of SSL in the skin cancer context. Figure 2 demonstrates the framework that will be executed in this work.

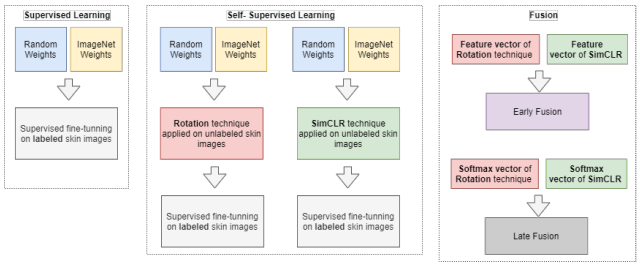


Fig. 2. Overview of the evaluated pipelines.

There has been developed an experimental framework to:

(i) investigate the impact of SSL on the training and generalization of a CNN for skin lesion diagnosis, and demonstrate that even with a small dataset there are benefits in using SSL. In order to better compare the impact of SSL the model was trained using two initialization techniques: random and ImageNet weights.
(ii) compare two different SSL approaches, one based on geometric distortion (Rotation) and another on contrastive learning (SimCLR).
(iii) provide a qualitative assessment of the impact of the different pre-training strategies, using explainability approaches (Grad-CAM [17] and LIME [18]).
(iv) demonstrate the complementarity of the features learned by the SSL strategies and the benefits of combining them.

This thesis will use two SSL techniques, which are believed to have a good performance on the skin image classification problem: Rotation [19] and the SimCLR [20].

*1) **Rotation**:* Rotation is a classification-based technique, where the network is trained to predict which rotation (0º, 90º, 180º, or 270º) has been applied to the image. Therefore, by predicting which rotation was applied to the input, the model is capable of extracting useful information from each image. The training pipeline starts with a small set of geometric transformations, which will be applied to the dataset. Secondly, the transformed images are fed to the model and the CNN is trained to identify which rotation was applied to the original image. As mentioned before, the set of geometric transformations defines the classification task, meaning that if there are four rotations then it is a 4-class classification problem. Figure 3 describes the proposed framework approach for the Rotation model.

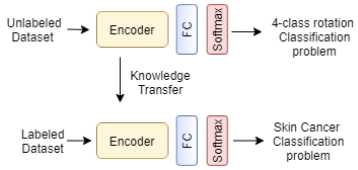


Fig. 3. Proposed framework using the Rotation technique. The last layers of the pretext task network are replaced by a FCL to output 8 classes.

*2) **SimCLR**:* SimCLR [20] is a SSL approach that applies the concept of contrastive learning to infer feature representations from the unlabeled dataset. Feature representations are learned by maximizing the agreement between differently augmented views of the same image via a contrastive loss, which will also accentuate the dissimilarity among different images. The key idea is when comparing the multiple images using the contrastive objective, the representations of corresponding views are 'attracted' to one another and the others are 'repelled'.

SimCLR can be divided into four main steps: 1) Random transformations are applied to the input, in order to obtain a pair of two augmented images, x_i and x_j; 2) Each augmented image within the pair is sent to an encoder, $f(.)$; 3) The output representations of the encoder, h_i and h_j, are then sent to a multi-layer perceptron (MLP), $g(.)$; 4) The contrastive loss is applied in the feature space, z_i and z_j, given by the MLP. Figure 4 describes the proposed framework approach for the SimCLR model.

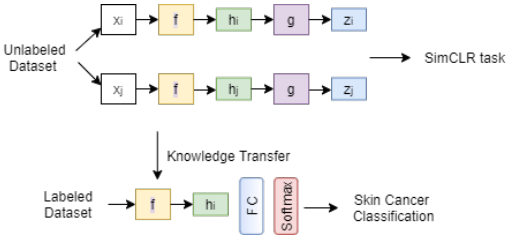


Fig. 4. Proposed framework using the SimCLR technique. The last layers of the pretext task network are replaced by a FCL to output 8 classes.

*3) **Fusion: Rotation and SimCLR**:* It will also be executed a set of experiments that will combine both SSL techniques: Rotation and SimCLR. Both SSL techniques force the network to learn different tasks, which results in two models that might learn different information. However, the question that arises is: 'Is the information of both techniques complementary?'.

The goal of conducting these tests is to improve the global performance of both methods, assuming that each model car-

ries different information about each skin lesion. To combine both Rotation and SimCLR techniques it will be used the early and late fusion approaches. Both techniques differ at the level of fusion: early fusion concatenates the models in a feature level, while late fusion fuses the models in the classification scores levels [21].

D. Feature Assessment

This section will explain two algorithms used to understand what a CNN sees to make a decision.

1) ***Grad-CAM****:* Gradient-weighted Class Activation Mapping [17], also known as Grad-CAM, uses the gradient information that the last convolutional layers of the CNNs have to determine the importance weights that each neuron has for the predicted class. Therefore, the main goal of Grad-CAM is to explore the spatial information preserved in the convolutional layers to better comprehend the parts of the input that contributed to the predicted decision. This method could explain activations in any layer of a deep network. However, it is mainly used in the last convolutional layers of the network since these layers have the best compromise between spatial information and high-level semantics.

2) ***LIME****:* Local Interpretable Model-agnostic Explanations [18], also known as LIME, is an explanation technique that intends to explain the predictions of a classifier by changing its input and understanding how its predictions are altered. To ensure that the explanation is interpretable, LIME modifies the original feature space and the interpretable representation. Therefore, this algorithm generates a new dataset containing perturbed samples and the corresponding predictions. In images, perturbing individual pixels do not make much sense, since many pixels contribute to one class. Hence, LIME creates variations in the images by first dividing the image into groups of pixels, known as 'super-pixels', and switches them on and off. Super-pixels are interconnected pixels that have similar textures and can be turned off by replacing each pixel with a gray color. Therefore, in images, the interpretable space is a binary vector indicating the presence or absence of a super-pixel. This means that to obtain the explanation of the prediction, the image is perturbed by hiding one or more super-pixels to get the corresponding prediction.

IV. Experimental Results

This chapter starts by introducing the dataset and metrics used to evaluate all the experiments. Afterwards, it presents a description and discussion of the experimental results performed during this thesis.

A. Dataset and Evaluation Metrics

All experiments were performed using the ISIC 2019 [22] [23] [24]. This dataset comprises a total of 8,238 images for testing and for training it contains 25,331 dermoscopy images with ground truth labels, divided into 8 lesions classes: Actinic keratosis (AKIEC), Basal cell carcinoma (BCC), Benign keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Nevus (NV), Squamous cell carcinoma (SCC) and Vascular (VASC). These labels are only used to train the classification models (recall Fig. 1).

In order to compare the different initialization approaches and assess their robustness, there was adopted a 5-time Monte Carlo sampling strategy, where the ISIC 2019 dataset was partitioned five times into training (70%) and validation (30%) sets. Based on this, there was performed the median and standard deviation of the following metrics: Confusion Matrix, Balanced Accuracy (BACC), Precision, F1-Score, Specificity, and Area under the curve (AUC).

B. Network Training and Computational Environment

The experimental framework was implemented using Tensorflow/Keras and one NVIDIA Tesla K80 GPU. However, for the complementary study with the use of more data, it was opted to use a laptop computer with the following specifications: Processor: AMD Ryzen Threadripper 3960X 24-core; Memory: 128 GB RAM; Graphics Processing Unit (GPU): NVIDIA GeForce RTX 3090. All models were trained for 60 epochs, using early stopping and the Adam optimizer [25]. The batch size was set to 32. For SSL, the losses are the categorical cross-entropy for the rotation task and for the SimCLR it was used the NT-Xent loss (with $\tau = 0.1$). For this task, the input image was transformed using horizontal flips, central crops and rotations (0, 90, 180 or, 270 degrees). The impacts of random color distribution and random Gaussian blur were also studied, however these experiments resulted in a lower performance of the model. Both tasks had an initial learning rate of $\eta = 10^{-4}$, however, the rotation had a reduction factor of 0.75 and the SimCLR an exponential decay of 0.96. To train the classifier, the weighted categorical cross-entropy loss was adopted, where the weights are set to the relative frequency of each class, in order to account for the unbalance. Here the learning rate was set to $\eta = 10^{-5}$, with a reduction factor of 0.75.

C. Comparison between the different initialization techniques

This section is divided into two parts: i) a quantitative analysis, where a comparison between the different approaches taking into consideration the selected evaluation metrics is made; ii) a qualitative analysis that used the Grad-CAM technique [17] to convey a more interpretable analysis of the impact of the various initialization strategies in the features learned by the model;

1) ***Quantitative Analysis****:* Table I summarizes the median and standard deviation of the scores obtained for the different initialization techniques. By looking at Table I it is possible to see that there are some benefits in using SSL when compared to the baseline supervised training. By looking at the baseline trained from scratch (row 1) and to both rows trained from scratch with SSL techniques (row 3 and 4) it is visible that both SSL techniques presented higher median and lower standard deviations. This proves that when comparing models trained from scratch there is a tendency to have **higher accuracy** and **more stability** (the standard deviation has a lower value) in the **models** that use **SSL**. By looking at the models trained using the ImageNet weights - the baseline (row 2) and to both

TABLE I. Application of the Monte Carlo Sampling with different initialization techniques:training the model from scratch or fine-tuning with ImageNet weights; application of two SSL techniques -Rotation and SimCLR.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Baseline	Scratch	46,82 ± 2,00	35,37 ± 3,84	37,24 ± 4,64	92,89 ± 0,55
	Imagenet	71,48 ± 1,82	65,14 ± 2,78	67,93 ± 1,75	96,04 ± 0,12
Scratch + SSL	Rotation	54,92 ± 1,15	40,54 ± 1,84	43,19 ± 2,04	93,39 ± 0,18
	SimCLR	52,54 ± 0,86	44,62 ± 1,39	47,53 ± 0,96	93,94 ± 0,18
Imagenet + SSL	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,70 ± 0,47	95,77 ± 0,05
	SimCLR	65,51 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18

models that used the SSL techniques (row 5 and 6) - it is visible that the latter two tend to have higher stability for all metrics (lower standard deviation) even though both had smaller or similar accuracy to the baseline. This proves that when comparing models trained with the ImageNet weights there is a tendency to **have more stability** in the models that use SSL.

Finally, looking at the SSL pre-trained models (row 2, 3, 4 and 5) and to the BACC column, it is possible to see that the **rotation technique has a higher accuracy** when compared to the model initialized with the SimCLR technique.

These results show that there **are benefits while using SSL** since there is **less variability in the performance of the classifier**. This proved that when combining TL with SSL the generalization problem that occurs when using TL is filtered. As mentioned before, TL uses natural images that have a different domain to the skin lesion ones. Therefore the network resulted from applying TL, will have neurons that remain loyal to the natural images. By applying SSL these neurons are 'corrected' and the obtained network generalizes better to the skin lesion images.

*2) **Qualitative Analysis**:* It was opted to execute a qualitative analysis, in order to understand what each model saw differently and what it learned in order to make the diagnostic decisions. Therefore, to analyze the differences between the learned representations for each initialization technique the Grad-CAM [17] was used. Figure 5 shows the Grad-CAM results for the different initialization techniques (fine-tuned with ImageNet weights).

Figure 5 proves that for the same input image all three models look at different parts of each lesion. Therefore, apart from having different performances, each model seems to learn different information about each class of lesion. The SimCLR pre-trained model tended to focus more in the parts of the lesion that presented higher contrast, while the Rotation looked more at the structure of each lesion. The ImageNet pre-trained model, was the least intuitive to interpret since its focus varied between lesion and skin. After, analyzing a set of different images it was possible to confirm that each method also had some limitations. The rotation had difficulties in detecting centered and symmetrical lesions, since each rotation of 90 degrees is similar, then the model does not learn useful information about this lesion. This limitation is visible in the fifth row of fig. 5. The SimCLR showed to be more precise

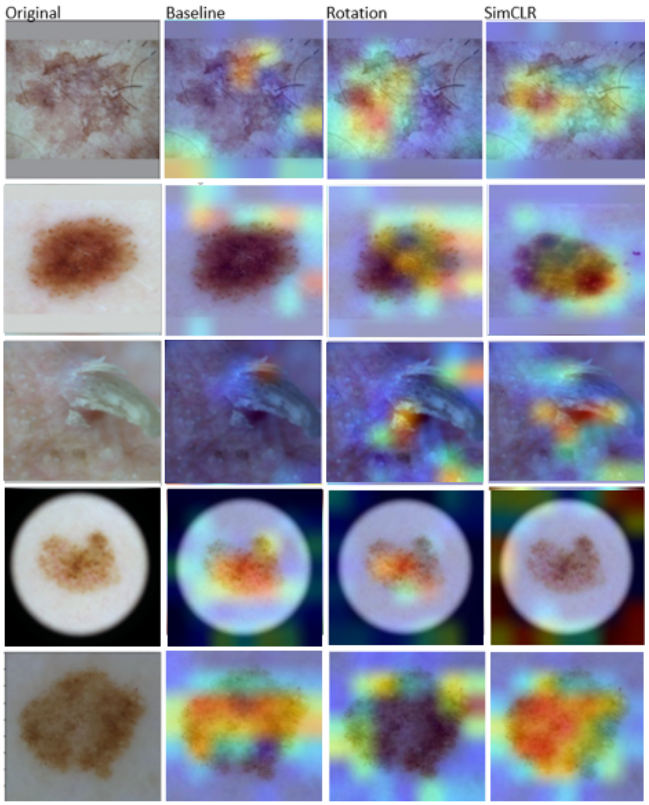


Fig. 5. Example of different lesion visualizations using the Grad-CAM algorithm (Baseline, Rotation and SimCLR).

in detecting the lesion. However as some images contained margins with high contrast (black borders), this method tended to focus more on the margins than the lesion (exemplified in the fourth row of fig. 5). Based on the qualitative results, the question that arose next was: Is the information learned by both SSL techniques complementary?

D. Fusion of SSL Approaches

As a consequence of the previous interrogation, two tests were performed that fused the models pre-trained with SSL. First, early fusion was used, this method fuses the different models in the feature space. Secondly, it was applied the late fusion technique, which fuses the models in the classification scores level (applied the mean strategy). The results were evaluated with a quantitative and qualitative analysis.

TABLE II. Application of the Monte Carlo Sampling with different initialization techniques (using ImageNet weights): application of two SSL techniques -Rotation and SimCLR- and fusion of both techniques.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Imagenet + SSL	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,70 ± 0,47	95,77 ± 0,05
	SimCLR	65,51 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
Fusion	Early Fusion	**73,78 ± 0,24**	**68,41 ± 4,13**	**70,99 ± 2,61**	**96,40 ± 0,36**
	Late Fusion (mean)	57,09 ± 2,19	50,28 ± 1,41	52,02 ± 1,08	94,24 ± 0,19

*1) **Quantitative Analysis**:* Table II presents the fusion results. It is possible to conclude that the **early fusion (row 3) had better results than any other model both in stability and accuracy**, proving that, in fact, the features of both models have complementary information. However, the late fusion (row 4) proved to have worse results, meaning that the features are complementary, but not learned classification models.

*2) **Qualitative Analysis**:* To analyze the learned representations of the fused model, the Lime algorithm [18] was used. It was opted to only analyze the model that used Early Fusion since it had better results.

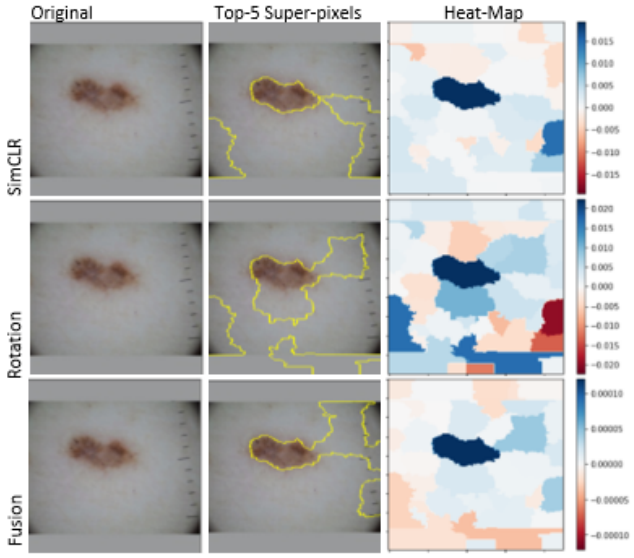


Fig. 6. Example of a lesion visualization using the Lime algorithm for each SSL pre-trained model (Rotation, SimCLR and Early Fusion).

Figure 6 shows the output obtained using the Lime algorithm for the different models. By analyzing this figure, it was possible to conclude that, for the same input image, both SSL pre-trained models look at different parts of each lesion (first and second rows). Additionally, it is also possible to verify that the model, resulted from early fusing the features of both SSL techniques, also looks at different aspects of the image and, combines the learned information from both models (last row). Looking at fig. 6 it is visible that the fused model was more precise in highlighting the skin lesion since the weights given for the Rotation and the SimCLR when combined resulted in higher importance in the lesion part.

As expected, this qualitative assessment proved that the fused model, apart from having higher performance, was also more accurate in detecting the different skin lesions. Therefore, this proved that the learned information of both SSL pre-trained models is, in fact, complementary. However, the question that arises is: 'Apart from being complementary is the combined information sufficient to avoid some of the limitations of each SSL technique? Figure 7 shows an example of a skin lesion where both the SimCLR or the Rotation pre-trained models had difficulties in detecting the skin lesion.

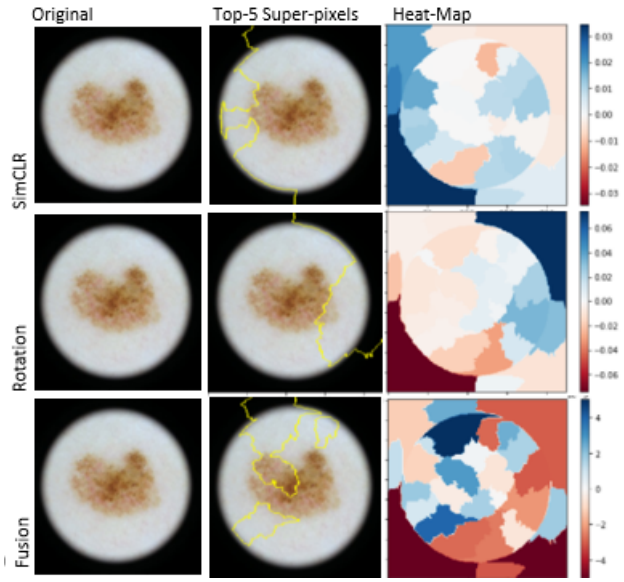


Fig. 7. Example of a lesion visualization using the Lime algorithm for each SSL pre-trained model (Rotation, SimCLR and Early Fusion).

Looking at figure 7 it is visible the lesion has less contrast than the margin and since it is quite symmetrical, both SSL pre-trained models had difficulties in detecting this lesion. However, when the features are combined the importance weights tended to highlight the lesion (visible in the last row). Therefore, by combining both models some of the limitations presented in both SSL pre-trained models could be avoided.

E. Further Quantitative Evaluation of all initialization techniques

To corroborate the conclusions made by analyzing table I and table II a boxplot was implemented. Figure 8 presents the boxplot containing all different initialized models (minus the late fusion since it had a worse performance).

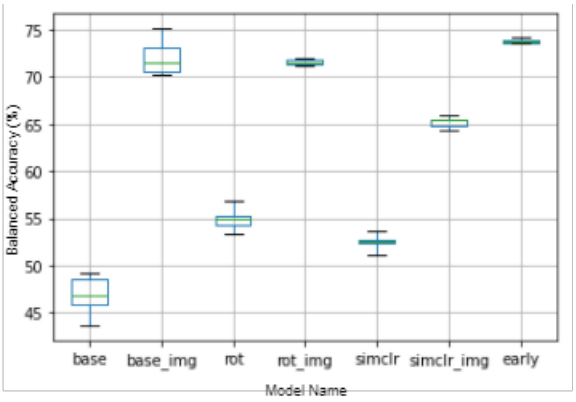


Fig. 8. Boxplot of the different implemented models. The green line represents the median and the box represents the middle 50% of all data points. Baseline models - 'base'; rotation - 'rot', early fusion - 'early' and the models fine-tuned with ImageNet weights end with 'img' in their name.

Looking at figure 8 it is possible to confirm that the model with the highest accuracy and stability is the early fusion. This model gathers both learned features from the Rotation and SimCLR pre-trained models and it confirms that this learned information is complementary. The Rotation (rot_img) model has similar accuracy as the baseline model pre-trained in ImageNet (base_img), however it is more stable. Both SimCLR (simclr and simclr_img) and Rotation (rot and rot_img) models have higher stability than the baseline (base and base_img), this is possible to confirm since the box is narrower for both self-supervised models.

1) ***State-of-the-Art comparison****:* SSL has been used in the skin image context. Both Li *et al.* [8] and Tajbakhsh *et al.* [5] applied SSL techniques with color-based pretext tasks to the segmentation of skin lesions. Kwasigroch *et al.* [3] applied two SSL techniques based on geometric distortion to the skin cancer classification task. The closest work to the one executed in this thesis is that of Chaves *et al.* [9], in which they assess five SSL contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and improving model accuracy.

Therefore, to better compare this thesis trained models with the state-of-the art works, the AUC score was implemented, the results are presented in table III. This table presents this thesis AUC score as well as both the Kwasigroch *et al.* [3] and the Chaves *et al.* [9] results. It is important to recall that all three works have been trained using different datasets with different purposes. The ISIC 2017 [26] task had the objective of differentiating two classes - malignant (MEL) and benign (NV and BKL). The ISIC 2020 [27] had the same purpose, but it included more lesions within each class: benign (NV, atypical melanocytic proliferation, café-au-lait macule, lentigo NOS, lentigo simplex, solar lentigo, lichenoid keratosis, and BKL) and malignant (MEL).

Looking at table III, it is possible to confirm that the results presented in this thesis have higher AUC score than the ones presented in the Kwasigroch *et al.* [3] work. In addition, looking at the scores obtained in the Chaves *et al.* [9] work, it is visible that this thesis best work, which is the early fusion model, presented a better performance than most models (Sup. Baseline, BYOL, InfoMin, and MoCo). However, the early fusion model showed to have lower score than both the SimCLR (-0.66%) and the SwAV (-0.36%) models. Analyzing the standard deviation it is possible to conclude that the results obtained in this thesis show even less variability than the ones presented in the Chaves *et al.* [9] work.

TABLE III. Evaluation of the different models using the AUC score.

Authors	Dataset	Technique	AUC (%)
Kwasigroch et al., 2020 [3]	ISIC 2017	Jigsaw [28]	83,4
		Rotation [19]	84,2
Chaves et al., 2021 [9]	ISIC 2020	BYOL [29]	94,6 ± 0,5
		InfoMin [30]	94,4 ± 0,5
		MoCo [31]	93,9 ± 0,7
		SimCLR	**95,6 ± 0,3**
		SwAV [32]	95,3 ± 0,6
Thesis work	ISIC 2019	Baseline	94,6 ± 0,3
		Rotation	94,7 ± 0,2
		SimCLR	92,9 ± 0,1
		Early Fusion	**94,9 ± 0,2**

F. Complementary Study: Study the impact of adding more data to the SSL pre-trained models

SSL is known to benefit from using more data. In the pre-training phase, this technique does not use labels, therefore the performance of the network increases with the variability of the available data. The more data, the more accurate the model can be to execute the intended SSL technique. It is also important to recall that depending on the level of difficulty of the task, the more it benefits from using more data. The impact that adding more data would have on the SSL pre-trained models was studied. It was opted to add 50% more data (using the ISIC 2020 dataset [27]). Table IV gathers both the evaluation metrics obtained from training the previous models using 50% more data and the results of table II.

Analyzing table IV it is possible to conclude that, in fact, the SimCLR task benefited from the use of more data. On the other hand, the Rotation technique had similar metrics to the previous training. This could be explained by the fact that this is a simpler task. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.

G. Final Evaluation in the Test Set

In order to verify how well the models obtained in this thesis generalized, it was opted to evaluate them using the test set provided by the ISIC 2019. This is an independent set without ground truth data and the evaluation of the models was performed on an online platform [33].

To compare the results obtained using the test dataset[1], the ISIC leaderboard [34] was analyzed. The classification in this challenge is based on the weighted accuracy of all classes (weighted average of the SE). It is important to recall that the test dataset contains a class unknown. However, in this

[1]The pre-processing of the test set instead of adding the most predominant color of the image, it was opted to add black margins, since it accelerated the pre-processing process.

TABLE IV. Application of the Monte Carlo Sampling using more 50% of unlabeled data.

SSL Dataset	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
ISIC 2019	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,7 ± 0,47	95,77 ± 0,05
	SimCLR	65,51 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
	Early Fusion	73,78 ± 0,24	68,41 ± 2,07	70,99 ± 2,61	96,40 ± 0,36
50% more data	Rotation	70,22 ± 0,98	62,89 ± 1,56	66,04 ± 0,96	95,73 ± 0,39
	SimCLR	**67,48** ± 0,58	**64,34** ± 6,05	**65,16** ± 3,69	**95,44** ± 0,63
	Early Fusion	**74,28** ± 0,58	**71,15** ± 1,57	**73,03** ± 0,96	**96,41** ± 0,15

thesis, it was opted to use the BACC score (without taking into account the class unknown) since the same importance is given to all the classes, even if they contain a different number of examples.

Table V contains the performance achieved by the different initialization models implemented in the validation and held-out test set for the best partition. The accuracy containing the class unknown is presented in the column 'Test w/ UNK class' of table V. However, the BACC without considering the unknown class is presented in the column 'Test'[2] of table V.

TABLE V. Evaluation of the different models using the test set.

SSL Dataset	Technique	BACC		
		Valid	Test w/ UNK	Test
ISIC 2019	Baseline	0,715	0,435	0,438
	Rotation	0,715	**0,454**	**0,471**
	SimCLR	0,656	0,417	*
	Early Fusion	0,738	0,424	0,452
50% more data	Rotation	0,712	**0,452**	**0,473**
	SimCLR	0,675	0,445	0,421
	Early Fusion	0,743	0,427	0,466

Analyzing table V it is possible to verify that both the SimCLR and the fused model increased their BACC score performance in the test set with the use of more data. This contributed to prove that the more data, the more accurate the model can be to execute the intended SSL technique depending on the level of difficulty of the task. Meaning that the Rotation technique showed little improvement since it is a simpler task than the SimCLR technique. It is also visible that the model with better results in the evaluation of the test set is the Rotation model. This could be explained by the fact that the pre-processing process of the dataset was made using padding of black margins, which can be a limitation of the SimCLR model. The model tended to focus more on the parts of higher contrast of the image, which in this case were the margins. Therefore, since the SimCLR had a worse performance the fusion of both models also had difficulties due to the higher contrast in the margins. Additionally, it is interesting to verify that both the Rotation and the Early Fusion models had a higher performance than the baseline model.

[2]The entries containing '*' were not presented in the top 200 of the online platform and, therefore, the SE score was not available. Meaning that the BACC could not be calculated.

V. Conclusions and Future Work

This thesis performed a robust assessment of the impact of SSL as a pre-training technique for skin cancer diagnosis. In particular, it performed a quantitative and qualitative analysis of the different pipelines. During this assessment, two SSL techniques were compared: Rotation and SimCLR. The experimental results show that there are benefits while using SSL. It was possible to observe that when applying these techniques, the classification CNN appeared to have more stability in its performance. It is beneficial to have models that are more stable since this means they are more trustworthy to apply to other data. Additionally, this proved that when combining transfer learning with SSL, the generalization problem that occurs when using TL is filtered. TL uses natural images that have a different domain to the skin lesion ones. Therefore the network resulted from applying transfer learning, will have neurons that remain loyal to the natural images. By applying SSL these neurons are 'corrected' and the obtained network generalizes better to the skin lesion images. This is believed to be the first work that provided a qualitative analysis of the features learned by the SSL strategies. This study led to the conclusion that each model learned different information from the data. Additionally, it was also possible to conclude that each SSL technique had some limitations: the Rotation had difficulties in detecting symmetrical lesions, while the SimCLR, as some images contained margins with higher contrast, tended sometimes to focus more on the margins than the lesion itself. In order to verify if the information learned by both SSL models was complementary, it was studied the combination of both techniques that resulted in the highest performance ($BACC = 73,780,24\%$). In addition, it was also possible to conclude that the model resulted from combining both SSL techniques overcame some limitations that each SSL model had individually.

As SSL is known to benefit from using more unlabeled data, it was also studied the impact of adding 50% more data to the SSL pre-trained models. It was possible to observe that depending on the level of difficulty of the task, the more the model benefits from using more data. Therefore, the SimCLR task benefited more from the increase of data, since this is a more complicated task when compared to the Rotation. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved. Finally, the pre-trained models were evaluated using the test set. This study reinforced the

conclusion that the SimCLR model trained using more data had higher capability to generalize to new data. Additionally, the Rotation and the Early fusion models have also shown to have higher performance than the baseline model even in the test set.

A. *Future Work*

The results obtained in this thesis highlighted the importance of using SSL techniques. However, there is room to improve the results. Therefore, some points can be highlighted regarding some topics that can be studied in future works: i) During this thesis for the SimCLR there were used three combinations of image transformations: horizontal flips, central crops and rotations (0, 90, 180, or 270 degrees). The impact of random color distribution and random gaussian blur were also evaluated, however, these experiments resulted in a lower performance of the model. In the future it could be interesting to try less 'aggressive' transformations such as normalizing each image color, which could result in better performance; ii) Instead of combining both models using early fusion, it could be interesting to train a single network to execute both SSL techniques simultaneously and, therefore, its final performance could have better results; iii) It could be interesting to try another SSL technique and combine it with the SimCLR and the Rotation technique using the Early Fusion method. Therefore, since the Rotation technique focus in the structure of the lesion and the SimCLR in the contrast, it could be beneficial to implement in the future a technique related to color such as the ColorMe [8] technique.

REFERENCES

[1] *Euromelanoma: Cancro da Pele, 2020*.

[2] Yasuhiro Fujisawa, Sae Inoue, and Yoshiyuki Nakamura, "The possibility of deep learning-based, computer-aided skin tumor classifiers," *Frontiers in medicine*, vol. 6, pp. 191, 2019.

[3] Arkadiusz Kwasigroch, Michał Grochowski, and Agnieszka Mikołajczyk, "Self-supervised learning to increase the performance of skin lesion classification," *Electronics*, vol. 9, no. 11, pp. 1930, 2020.

[4] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flavia Vasques Bittencourt, Sandra Avila, and Eduardo Valle, "Knowledge transfer for melanoma screening with deep learning," *2017 IEEE 14th ISBI, 2017*, pp. 297–300, Apr 2017.

[5] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *2019 IEEE 16th ISBI, 2019)*, 2019, pp. 1251–1255.

[6] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, pp. 101539, 2019.

[7] Carl Doersch and Andrew Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE ICCV*, 2017, pp. 2051–2060.

[8] Yuexiang Li, Jiawei Chen, and Yefeng Zheng, "A multi-task self-supervised learning framework for scopy images," *2020 IEEE 17th ISBI, 2020*, pp. 2005–2009, 04 2020.

[9] Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila, "An evaluation of self-supervised pre-training for skin-lesion analysis," *CoRR*, vol. abs/2106.09229, 2021.

[10] Ivars Namatevs, "Deep convolutional neural networks: Structure, feature extraction and training," *Information Technology and Management Science (Sciendo)*, vol. 20, no. 1, 12 2017.

[11] Gilu K Abraham, VS Jayanthi, and Preethi Bhaskaran, "Convolutional neural network for biomedical applications," in *Computational Intelligence and Its Applications in Healthcare*, pp. 145–156. Elsevier, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE CVPR*, 2016, pp. 770–778.

[13] S. Chan, Vid Reddy, B. Myers, Q. Thibodeaux, N. Brownstone, and Wilson Liao, "Machine learning in dermatology: Current applications, opportunities, and limitations," *Dermatology and Therapy*, vol. 10, 04 2020.

[14] Mark Ryan M. Talabis, Robert McPherson, I. Miyamoto, Jason L. Martin, and D. Kaye, *Chapter 1 - Analytics Defined*, pp. 1 – 12, Syngress, Boston, 2015.

[15] Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and Pengtao Xie, "Transfer learning or self-supervised learning? a tale of two pretraining paradigms," *ArXiv*, vol. abs/2007.04234, 2020.

[16] Graham D Finlayson and Elisabetta Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2004, vol. 2004, pp. 37–41.

[17] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct 2019.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Apr. 2018.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[21] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan, "Learn to combine modalities in multimodal deep learning," *CoRR*, vol. abs/1805.11730, 2018.

[22] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 08 2018.

[23] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern, "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 ISBI, 2018*, 2018, pp. 168–172.

[24] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy, "Bcn20000: Dermoscopic lesions in the wild," *ArXiv*, vol. abs/1908.02288, 2019.

[25] Mohammed Alom, "Adam optimization algorithm," 06 2021.

[26] N. Codella, D. G., M. Emre C., B. H., M. Marchetti, S. D., A. K., K. L., N. Kumar M., H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the isic," 10 2017.

[27] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.

[28] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," *In ECCV, Germany*, pp. 65–84, 2016.

[29] J. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Pires, Z. Guo, M. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Neural Information Processing Systems, Montréal, Canada. hal-02869787v2*. Elsevier, 2020.

[30] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, "What makes for good views for contrastive learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 6827–6839.

[31] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF CVPR*, 2020, pp. 9726–9735.

[32] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *CoRR*, vol. abs/2006.09882, 2020.

[33] "Isic challenge," .

[34] "Isic archive, 2020," .