RESET: A RESIDUAL SET-TRANSFORMER APPROACH TO TACKLE THE UGLY-DUCKLING SIGN IN MELANOMA DETECTION

Jules Collenne, Rabah Iguernaissi, Séverine Dubuisson, Djamal Merad

Aix Marseille Université, CNRS, LIS
Marseille, France

ABSTRACT

The dermatological concept of the Ugly-Duckling Sign (UDS) emphasizes the importance of comparing skin lesions within the same patient for enhanced diagnostic accuracy in melanoma detection, stating that atypical lesions are more likely to be cancers. However this concept is still underutilized in research, as most work on melanoma detection rely on classification ConvNets which lack the capacity to compare images together. Addressing this research gap, we introduce ReSeT (Residual Set-Transformer), a framework designed to compare skin lesions within patients during prediction. ReSeT comprises an encoder that takes individual images as input to generate embeddings, and a Set-Transformer with a residual prediction layer that compares these embeddings while predicting. We demonstrate that our architecture ReSeT significantly enhances performance compared to ConvNets and we highlighting the necessity of residual connections in the context of multi-output Transformers. We also observe that self-supervised encoders are able to generate embeddings of comparable quality to those of supervised models, showing their robustness and impact on image comparison tasks.

Index Terms— Skin lesions, Set-Transformers, Self-supervised learning, Machine Learning, Ugly-Duckling Sign

1. INTRODUCTION

Melanoma is a prevalent and potentially life-threatening skin cancer occurring worldwide. The first step toward its diagnosis is to analyze the skin's surface and identify potentially dangerous skin lesions. Initially performed by clinicians, an increasing body of research has focused on automating diagnosis using artificial intelligence. While most studies rely on conventional ConvNets, which have already demonstrated good results [1], they often lack incorporation of crucial dermatological concepts. The integration of these concepts into automated diagnostic models shows promise in improving both diagnostic performance and the clinical acceptability of such models in clinical settings. Within the extensive landscape of dermatological concepts documented in medical literature, the Ugly-Duckling Sign (UDS) [2] emerges as a distinctive and insightful perspective. UDS posits that among all the skin lesions on a patient's skin, those that appear the most different from the surrounding lesions are more likely to be indicative of malignancy. While it represents a promising research direction, there are still few works related to this concept [3, 4, 5]. Integrating patient-level comparisons could significantly enhance diagnostic accuracy and the depth of clinical insights when deploying computer-aided models within clinical settings. Classical Transformers [6], while showing promising initial results [7], rely heavily on token order for predictions and may not be optimized for data organized as sets. In our patient-focused context, treating skin lesions within a patient as a set, rather than an ordered sequence, might prove beneficial for the classification task and the efficiency of our models. Contrary to previous work [7], which leverages a wide range of techniques to compare skin lesions, we aimed for our architecture to remain simple and adaptable, focusing solely on the task of classifying images while minimizing computing resources. Following this idea, our framework (illustrated in Figure 1) rely on an encoder in conjunction with a Set-Transformer [8] to facilitate diagnostic predictions. The encoder, either trainable in a supervised or self-supervised manner, generates embeddings of the images, which can be used for various subsequent tasks. In our case, we focus on the classification task in order to detect melanomas. We emphasize the use of self-supervised models as encoder, as these models, by abstaining from the use of ground truth labels, ensure that the extracted embeddings remain unbiased towards the original classification task. This is particularly crucial when comparing lesions to obtain a more diverse representation of the images, with embeddings representing features that would not necessarily be learnt by a supervised model. Subsequently, a Set-Transformer is employed to compare these embeddings and drive diagnostic predictions. Their permutation-equivariant design ensures that predictions remain consistent irrespective of lesion order, which is the case in our task.

Thanks to the French National Research Agency for the funding of the DIAMELEX project (ANR-20-CE45-0026).

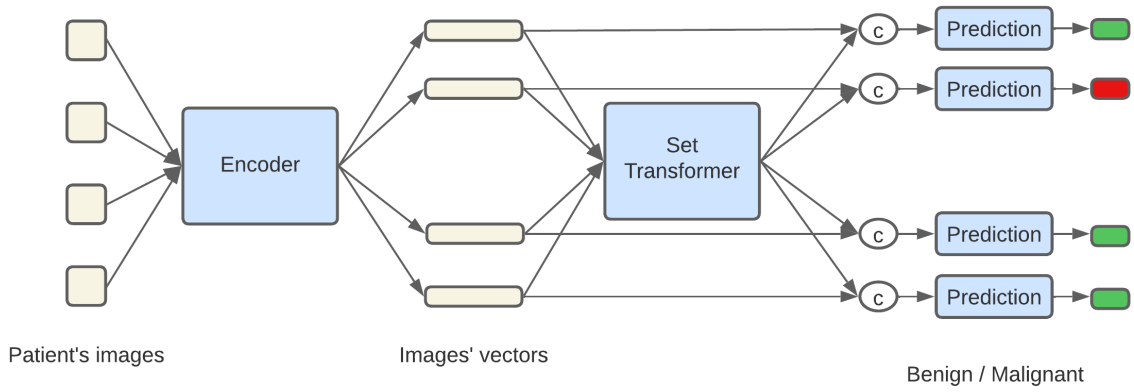

Fig. 1. **Framework Diagram.** Images are processed individually by the encoder, whereas the Set-Transformer operates collectively on all vectors for comparison. The Set-Transformer's outputs are then concatenated with the input vectors during predictions.

2. METHOD

2.1. Dataset

The dataset employed in this study is the ISIC 2020 dataset [9], one of the most important dataset on skin lesions. It contains a substantial collection of skin lesion images that uniquely includes patient identifiers. Images come from multiple dermatology departments and have been acquired through the use of epiluminescence microscopy. With over 33,000 annotated images, it is one of the most extensive dataset for skin lesion analysis. Notably, the patient identifiers enable us to link lesions from the same patient, a critical aspect of our approach. While being a very rich dataset, the data are heavily imbalanced, with nevus representing more than 98% of the images, and melanoma being part of only 2% of the images (see Table 1). To deal with such imbalance, our supervised models (classification CNN and Set-Transformers) have a weight in favour of melanoma corresponding to 0.98, and a weight 0.02 for nevus class. In preparing the dataset, we conducted several data preprocessing steps. Apart from resizing all images to a uniform size of 300×300 pixels, we also applied data normalization to ensure consistent image quality and format. Additionally, we employed data augmentation techniques depending on the architecture used. We kept the default techniques for each self-supervised model, while the CNN used basics transformations such as random resized crops and random flips. The dataset was randomly divided into three subsets: training, validation, and test set. The training set encompasses 75% of the patients, the validation and test sets each contain 15% of the patients, providing a robust basis for model evaluation and generalization testing. It is worth highlighting that our dataset splitting was performed at the patient level rather than directly at the image level. This approach is essential as each training step requires access to all images from the same patient.

	Train set	Val set	Test set
Number of patient	1439	308	309
Nevus	23081	5093	4368
Melanoma	409	105	70

Table 1. Distribution of images in the different sets.

2.2. ReSeT

ReSeT consists of two primary modules: an encoder and a Set-Transformer, trained independently. Firstly, the encoder is trained to derive image embeddings, essential for subsequent operations within the Set-Transformer. Then the Set-Transformer is trained on the classification task, taking as input the generated embeddings. We use the same train, validation, and test sets as the encoder. The Set-Transformer produces an output for each input, mirroring the patient's lesion count, with each output corresponding to the related input image. Additionally, we augment the original Set-Transformer architecture by introducing a final residual layer (skip-connection) that links the embeddings to the corresponding output. This modification involves concatenating these outputs with their respective input vectors, rather than relying solely on the Set-Transformer's outputs. Finally, a single dense layer is used to make the predictions. Residual layers ensure that the network's final layer assimilates both the original vector and contextual information from the Set-Transformer. While the encoder learns to map images in the latent space, the Set-Transformer learns to make predictions not only using the individual images but also leveraging addi-

tional information related to all other lesions. This approach allows the model to either utilize contextual information or simply rely on the image embedding, depending on the patient, as some patients may not have many skin lesions. Moreover, this method addresses concerns regarding vanishing gradients while maintaining a balance between individual image traits and collective set features in predictions.

2.3. Encoder

We utilized ResNet-50 [10] as the base model for the architectures responsible for encoding images. Since the original architecture offer a latent space with more than 1,000 features, which is too high in our context, we added a dense layer with an output size of 200, aiming for concise, insightful embeddings rather than high-dimensional vectors. In our supervised training setup, we trained the ResNet-50 model on a classification task distinguishing between melanoma and nevus on 200 epochs, using the Adam optimizer, a learning rate set at 0.001, and a batch size of 64. The model selection was based on the Area Under the ROC Curve (AUC) from the validation set. Post-training, we extracted image embeddings from the penultimate layer of size 200. We also explored our framework's performance by employing self-supervised models as encoders. Leveraging unlabeled data through self-supervised learning enables the capture of underlying data structures beyond mere classification labels. Most of these self-supervised allow the generation of a latent space where proximity indicates image similarity, which offer advantages in classification tasks as well as visualization tasks in clinical settings. We compared current major self-supervised architectures such as BYOL [11], MoCo [12], SimCLR [13], SwaV [14], and SimSiam [15]. Each model has an unique method of mapping images into the latent space, which could potentially provide diverse and unique representation of the images.

BYOL [11] uniquely avoids the use of negative pairs and instead relies on two neural networks that learn from each other by minimizing the distance between their representations of two augmented views of the same image.

MoCo [12] stands out for its use of a dynamic dictionary of encoded representations and a momentum-updated encoder to maintain consistency over time, facilitating effective contrastive learning even with a large number of negative samples.

SimCLR [13] distinguishes itself with its simple yet effective framework that uses a large batch size and enhanced data augmentation to learn powerful representations by maximizing agreement between differently augmented views of the same data.

SwAV [14] uniquely clusters the data while enforcing consistency between cluster assignments of different augmented views of the same image, using a technique called "swapped prediction" that improves representation by comparing cluster assignments instead of direct feature vectors.

SimSiam [15] is unique in its approach of using a siamese network architecture with no negative pairs and preventing collapse through stop-gradient operations, focusing solely on similarity maximization between two augmented views of the same image.

Each model underwent 200 epochs of training with a batch size of 128, with default learning rates specified in the related papers, and maintaining a standardized output size of 200 units. These models were initialized randomly, bypassing the use of ImageNet weights, thus assessing the framework's ability to discern melanoma from nevus without prior knowledge. Apart from these, all other hyperparameters were kept the same as described in their related article.

2.4. Set-Transformer

Set-Transformers are well-suited for our task of comparing skin lesions due to their efficiency in handling unordered data sets. The original paper [8] introduces operations like SAB (Set-Attention Block), ISAB (Induced SAB), Multihead Attention Block (MAB), and Pooling by Multihead Attention (PMA), enabling permutation invariance and optimization of computational complexity by using inducing points, thereby efficiently processing unordered data while minimizing computational costs. Our utilized Set-Transformer model adopts an encoder-decoder structure, similar to original implementation. The encoder comprises two ISAB blocks, while the decoder includes a PMA layer, two SAB blocks, and a final dense layer. We set the hidden dimension to a size of 128 and set 8 attention heads. During training, the Set-Transformer processes each patient's image list (represented as vectors), generating context information vectors of size 20. These outputs are concatenated with the original input vectors and fed into the final prediction layer. With a contextual output of size 20, we ensure that the generated vectors represent useful information for the prediction. Training spanned 200 epochs, using cross-entropy loss, SGD optimizer and a learning rate of 0.001. To prevent overfitting, training stopped after 15 epochs without validation AUC improvement. Each training step involved 64 patients, with 20 random vectors selected per patient, surpassing the average of 16 lesions per patient found in the ISIC 2020 dataset to ensure a more comprehensive representation. For patients with fewer lesions, zero vectors were used as padding. During validation and testing, patients with excess images were processed multiple times through the network for prediction. Leveraging Set-Transformers for skin lesion comparison introduces an innovative approach, harnessing the model's ability to process unordered sets and derive meaningful insights from images of skin lesions.

3. RESULTS

In this section, we conduct an extensive analysis of the outcomes derived from our classification task.

Supervised Architecture	AUC	Balanced accuracy	Sensitivity	Specificity
Classification CNN	0.905 ± 0.01	0.707 ± 0.01	0.775 ± 0.06	0.705 ± 0.06
Set-Transformer (without residual)	0.801 ± 0.01	0.690 ± 0.02	0.798 ± 0.09	0.583 ± 0.12
ReSeT	$\mathbf{0.924 \pm 0.01}$	$\mathbf{0.831 \pm 0.02}$	$\mathbf{0.874 \pm 0.05}$	$\mathbf{0.788 \pm 0.09}$

Table 2. **Test Set Classification Results.** The comparison includes a basic classification CNN, a Set-Transformer trained on the CNN's features without residuals, and the performance of ReSeT on the CNN's features.

Self-supervised Architecture	AUC	Balanced accuracy	Sensitivity	Specificity
BYOL	0.681 ± 0.01	0.635 ± 0.02	0.694 ± 0.04	0.576 ± 0.04
MoCo	$\mathbf{0.768 \pm 0.02}$	$\mathbf{0.708 \pm 0.02}$	0.730 ± 0.09	$\mathbf{0.686 \pm 0.06}$
SimCLR	0.764 ± 0.01	0.683 ± 0.02	0.780 ± 0.02	0.587 ± 0.02
SimSiam	0.748 ± 0.01	0.702 ± 0.01	$\mathbf{0.849 \pm 0.01}$	0.555 ± 0.02
SwaV	0.617 ± 0.01	0.569 ± 0.03	0.602 ± 0.16	0.536 ± 0.12

Table 3. **Classification results on self-supervised encoders.** Results of the classification task on the test set using self-supervised architectures for CNN training.

For each architecture, we selected the model with highest AUC on the validation set, and utilized a grid search methodology on the validation set to find the optimal classification threshold used for the computation of balanced accuracy, sensitivity, and specificity metrics. We kept the most suitable threshold for generating predictions from the probabilities. These optimized models, along with their respective thresholds, were subsequently utilized for the final prediction on the test set. Moreover, to comprehensively evaluate the performance of the architectures, we trained five distinct Set-Transformers for each architecture and computed their average and standard deviation. This method enabled us to analyze and compare the stability and average performance across architectures for each metric.

3.1. Supervised encoder

We begin by evaluating the performance of the supervised approach. This involves training a classification ConvNet on the binary task of detecting melanoma and nevus and use it as the encoder of the ReSeT architecture. Results are detailed in Table 2. The conventional classification CNN performed well, achieving an AUC (Area Under the Curve) of 0.905, aligning with expected performance for this task. Interestingly, feeding features directly into a classical Set-Transformer led to a decrease in performance, yielding an AUC of 0.801. This decline might arise from challenges in associating each output with its corresponding image input. However, our model ReSeT, which integrates skip-connections with the CNN's features, demonstrated a significant performance enhancement, achieving an AUC of 0.924. This result show the importance of using residual layer in this context, making the Set-Transformer learn how to predict each image along with the additional contextual information.

These findings highlight the potential of utilizing Set-Transformers in dermatological diagnosis, presenting a promising path for clinicians and researchers. The superior performance of the combined approach highlights the limitations of employing classical Transformers and Set-Transformers directly without tailoring them to the specific task at hand.

3.2. Self-supervised encoders

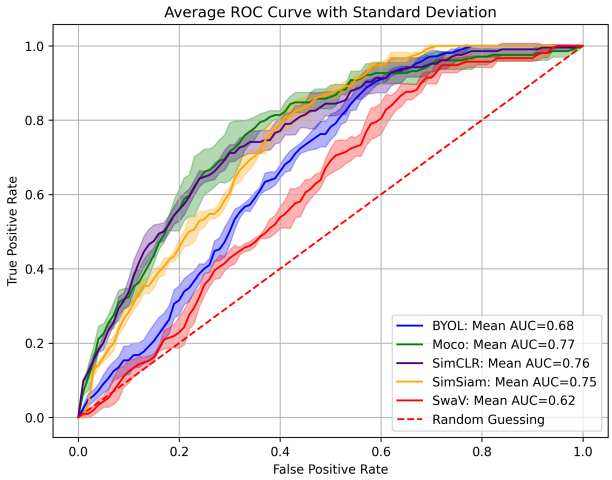

Fig. 2. Average ROC curves of the Set-Transformer trained on different encoders' features.

The outcomes derived from utilizing features extracted by self-supervised models are summarized in Table 3, accompanied by the corresponding receiver operating characteristic (ROC) curves depicted in Figure 2. Beyond all self-supervised architectures, MoCo consistently outperform other models, exhibiting an AUC of 0.768 and a balanced accuracy of 0.708, which correlate with its performances in other tasks in the literature. Moreover, it is very likely that the use of the dictionary lookup within MoCo enables the

model to grasp the different subcategories of skin lesions that exist in dermatology, leading to better performances. In contrast, alternative models show slightly lower performance, with their AUC ranging between 0.62 and 0.76. Results of self-supervised encoders are undoubtedly lower than supervised ones; however, it is notable that these architectures still achieve decent accuracy despite generating embeddings without access to any annotations. These results underscore the robustness and efficacy of MoCo's features, showcasing the potential of self-supervised approaches in generating useful features for image comparison. It is noteworthy that these encoders had no access to ground truth labels during training, emphasizing the significant achievement of these results.

4. CONCLUSION

We have introduced a novel framework for melanoma detection that harnesses the power of Set-Transformers alongside skip-connections and supervised and self-supervised encoders. Our approach has proven effective in efficiently computing robust features and identifying melanomas, resulting in a notable increase in classification model accuracy. These findings are of significant importance in the field of dermatology and computer-aided diagnosis, paving the way for further research on the Ugly-Duckling Sign. Coupled with recent advances in screening techniques and full-body imaging, research on image comparison is likely to have a significant positive impact on clinical settings, providing dermatologists with more insights when diagnosing skin lesions.

The features acquired through self-supervised models have the potential to significantly enhance prediction interpretability and enable thorough visualization of latent spaces, eliminating the need for the original classification task. These capabilities can empower dermatologists in their clinical practice, including visualization techniques that could be applied in clinical settings. We leave this path of research for further work, where some methods have already managed to obtain good results on images of faces [16].

Moreover, our approach exhibits broad applicability beyond dermatology, extending to a wide range of diseases and scenarios that require comparing multiple instances and predicting them. Its adaptability in cases where obtaining annotations is challenging or prone to noise positions it as a versatile and potent tool within the broader medical and healthcare domains.

5. ACKNOWLEDGMENTS

Thanks to the French National Research Agency for the funding of the DIAMELEX project (ANR-20-CE45-0026).

6. REFERENCES

[1] Q. Ha, B. Liu, and F. Liu, "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge," *arXiv:2010.05351*, 2020.

[2] J.-J. Grob and J.-J. Bonerandi, "The 'Ugly Duckling' Sign: Identification of the Common Characteristics of Nevi in an Individual as a Basis for Melanoma Screening," *Archives of Dermatology*, vol. 134, no. 1, pp. 103–104, 01 1998.

[3] J. J. Grob, Y. Wazaefi, Y. Bruneu, C. Gaudy-Marqueste, S. Monestier, L. Thomas, M.-F. Avril, R. Triller, G. Pellacani, J. Malvehy, and B. Fertil, "Diagnosis of melanoma: Importance of comparative analysis and "ugly duckling" sign.," *Journal of Clinical Oncology*, vol. 30, no. 15_suppl, pp. 8578–8578, 2012.

[4] C. Gaudy-Marqueste, Y. Wazaefi, Y. Bruneu, R. Triller, L. Thomas, G. Pellacani, J. Malvehy, M.-F. Avril, S. Monestier, M.-A. Richard, B. Fertil, and J.-J. Grob, "Ugly Duckling Sign as a Major Factor of Efficiency in Melanoma Detection," *JAMA Dermatology*, vol. 153, no. 4, pp. 279–284, 04 2017.

[5] J. Collenne, R. Iguernaissi, S. Dubuisson, and D. Merad, "Enhancing anomaly detection in melanoma diagnosis through self-supervised training and lesion comparison," in *Machine Learning in Medical Imaging*, Cham, 2024, pp. 155–163, Springer Nature Switzerland.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[7] Z. Yu, V. Mar, A. Eriksson, S. Chandra, P. Bonnington, L. Zhang, and Z. Ge, "End-to-end ugly duckling sign detection for melanoma identification with transformers," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Cham, 2021, pp. 176–184, Springer International Publishing.

[8] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds. 09–15 Jun

2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, PMLR.

[9] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and P. Soyer, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific Data*, vol. 8, pp. 34, 2021.

[10] 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 (CVPR)*, 2016.

[11] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds. 13–18 Jul 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*. 2020, vol. 33, pp. 9912–9924, Curran Associates, Inc.

[15] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15750–15758.

[16] X. Zhu, C. Xu, and D. Tao, "Where and what? examining interpretable disentangled representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5861–5870.