SSL based encoder pre-training for segmenting a heterogeneous chronic wound image database with few annotations

Guillaume Picaud[1], Marc Chaumont[1,2], Gérard Subsol[1], and Luc Téot[3]

[1]LIRMM, équipe ICAR, Univ. Montpellier, CNRS, France
[2]Univ. Nîmes Place Gabriel Péri, France
[3]Cicat-Occitanie, Montpellier, France
{guillaume.picaud, marc.chaumont, gerard.subsol}@lirmm.fr
l-teot@chu-montpellier.fr

Abstract. Segmentation is crucial in medical imaging, but obtaining a sufficient quantity of annotated data is challenging, limiting the development of high-performing deep learning models. Self-supervised learning (SSL) strategies offer a promising solution to address this lack of annotation. One such strategy, Dinov2 for Distillation with NO labels, enabled the creation of the vast LVD-142M database and the training of encoders, whose weights are now freely accessible. However, clinical images may not be well represented in LVD-142M. Thus, in the context of scarce annotated clinical data, we evaluate the benefits of a generic encoder pre-trained with DINO on LVD-142M versus a lighter one. We also explore the effect of SSL DINO pre-training strategy directly on the target dataset. We measure the impact of available label quantity on segmentation performances. Results show, in a context with few annotated images, specific and lightweight encoder can outperform generically pre-trained DINO one. Furthermore, DINO SSL pre-training on specific dataset is beneficial for small encoder.

Keywords: Self supervised learning · Dinov2 · segmentation · chronic wounds

1 Introduction

A chronic wound is a skin lesion that is not entirely healed after 4 or 6 weeks after it appears. Various factors can contribute to their occurrence in at-risk populations such as elderly, diabetic, and disabled people. Chronic wounds constitute a huge socio-economic issue with severe consequences for patients, ranging from amputation to death. In 2011, the French health insurance system estimated the cost of managing pressure sores and ulcers at home to be over one billion euros. Their prevalence is rising steadily [10], in particular, due to population aging. Teleconsultation is worldwide growing and enables to put in touch local paramedical teams with experienced physicians which improves general patient healthcare

inside the territory and fights against medical desert consequences. We can illustrate a teleconsultation scenario as follows : a care team is overwhelmed by the management of a polypathological patient's chronic wound that is stagnating or worsening. They decide to contact the chronic wounds expert association. After a first presentation of the case to assess the severity of the situation, a teleconsultation is scheduled. The expert will then hold a videoconference with the patient and his paramedical team to give precious advice regarding the diagnosis and prescribed care. In this way, Cicat-Occitanie experts have developed over 20 years of experience a considerable database of over 133,000 chronic wound photographic images taken by nurses with their personal smartphones directly at patients' homes. Those pictures are associated with valuable clinical information. Some examples are shown in Figure 1. This database is a precious but under-exploited resource, due to the lack of standardization of the acquisition protocol and image annotations such as chronic wound masks. Indeed, wound segmentation could the help medical team to better assess the effectiveness of the current treatment. However, manual tracing is a time-consuming task that can be challenging in certain cases even for experts. This results in significant discrepancies from inter and intra-annotator measurements. Furthermore, this expert database is characterized by the diversity of acquisition devices (smartphones used), scenes and conditions (different patient homes, variations in lighting, background, picture shooting, distance between smartphone and wound), and also wound type and location.

Deep learning methods are an efficient approach for segmenting chronic wounds. The Diabetic Foot Ulcer Challenge (DFUC)[1] makes available annotated database images but they were acquired in hospital conditions taken by experts. In 2022, organizers released two thousand images with their segmentation masks. This competition highlighted the performance of the HardNet-DFUS model [9]. However, the absence of similar initiatives for a heterogeneous database limits the development of deep learning approaches that cope with the actual diversity of clinical cases.

To overcome this issue, self-supervised learning (SSL) may appear as a promising lead. SSL enables neural networks to learn efficient image features without requiring human supervision. In particular, "Distillation with NO labels" methodology [1, 11], (DINO) uses the massive LVD-142M database on which Vision Transformer, ViT encoder [5] was trained to extract generic and discriminative feature in a very efficient way. However, clinical images are not specifically represented by this generic database. Thus, a pre-training on LVD-142M is sub-optimal.

This article aims to explore, in the clinical context of chronic wound segmentation with few annotations, the performance of a state-of-the-art generic encoder (pre-trained with DINO on LVD-142M) versus a lighter one, randomly initialized HarDNet-DFUS encoder. We also investigate the effect of SSL pre-training on target data before the final segmentation task. Finally, we measure the impact of the amount of available data for the final segmentation task with the various pre-training scenarios.

[1] https://dfuc2022.grand-challenge.org/

Fig. 1. Four pictures picked from the expert database illustrate the diversity of acquisition conditions, the chronic wound localization, and also the type.

2 Related work

SSL is an approach where an encoder is trained during a so-called pretext task in which, instead of using human annotations, relies on pseudo-labels automatically generated from the data itself [12, 16], as long as a huge quantity of data is available. Once pre-training is complete, the fine-tuned encoder's weights are used as initialization for the downstream task which most of the time is a supervised one. In this article, we are particularly interested in the discriminative approach illustrated by the DINO pretext task because of its state-of-the-art performances.

DINO uses two encoders of identical architecture, one called the student and the other the teacher. A multi-crop strategy [2] is applied to the input image: two "global" views, covering at least half of the original image, and n "local" views, with a surface area smaller than 50%, are produced. The teacher encoder will receive the two "global" crops while the student will see all of them. Each of these views is independently augmented using spatial and color transformations. During training, both encoders produce a representation of these views which will be passed to their respective projection heads, a series of linear layers (MLP). Generated feature maps are then compared using cross-entropy, and the student encoder's weights are updated through gradient backpropagation. The weights of the teacher are updated via an exponential moving average from those of the student.

Using the carefully curated LVD-142M database, SSL DINO enabled the training of ViT encoder at different scales (small 21 M, large 307 M, giant 1100 M parameters), see[2]. However, transformers remain memory-costly and compu-

[2] https://github.com/facebookresearch/dinov2

tational resource-intensive architectures. They manage to outperform convolutional approaches only when the databases are very large.
Hardnet [4], short for "Harmonic Densely Connected Network", is an enhanced convolutional architecture based on DenseNet [7], aiming to reduce inference time without compromising encoder performance. To achieve this, the number and position of residual connections within the convolutional blocks have been adjusted. Specific research on colon polyp segmentation task led to new contribution [6] in which the use of HarDNet encoder, combined with modified Cascade Partial Decoder decoder [14] achieved state-of-the-art performances in various colonoscopy polyp database in 2021.
The winner of the DFUC2022 competition used and refined module encoder HarDBlk [8]. The encoder was connected to a segmentation decoder called Lawin for Large window attention [15]. This encoder-decoder architecture is named HarDNet-DFUS [9] and is particularly promising for chronic wound analysis.

3 Experiences

3.1 Image Database preparation

The expert dataset gathers over 133,000 images of various types of chronic wounds (pressure ulcers, diabetic foot ulcers, etc.). This dataset has been automatically labeled with bounding boxes with the help of a Faster R-CNN [13] updated with the use of Deformable convolution [17] and previously trained as a wound detector using database [3] from the DFUC2020 competition[3]. Only images with a single predicted wound were retained, totaling 89,127 images. 400 images were then randomly selected for manual segmentation annotation by two experts using *labelme* annotation tool[4]. Those 400 annotated images from database B_2. Remaining images form database B_1.
For SSL pre-training, B_1 is divided into three categories: training, validation, and testing, with respective data ratios of 70%, 20%, and 10%.
For supervised segmentation fine-tuning, B_2 images are divided into five splits with repartition of 70%, 20%, 10% of B_2. There are no shared images between the five test sets. In those versions, we remove a certain number of randomly chosen training images, while the validation and test sets remain unchanged. As a result, we obtain three groups of five splits where the quantity of training data is respectively 280, 140, and 70. Thus, we can assess the impact of the training data quantity for fine-tuning. Five splits analysis is also interesting because it enables us to test any model on 200 different images in total.

3.2 Training scenarios

We selected ViT**l**14_reg (307 M) and ViT**s**14_reg (21 M) configurations, with initial weights from the Dinov2 article [11]. These two encoders are compared

[3] https://dfu2020.grand-challenge.org/
[4] https://github.com/labelmeai/labelme

with the one from HarDNet-DFUS (3 M), initialized randomly. Figure 2 summarizes the training scenarios. Encoders can either be pre-trained using the SSL DINO method on B_1 or undergo no pretext task. During the final task, encoder weights are either frozen or fine-tuned through a weight unfreezing strategy. Due to computational constraints, the weights of ViTl14_reg encoder are not optimized in this article. Also, fine-tuning with freezing strategy on HarDNet-DFUS without SSL pre-training (i.e. from scratch) is avoided here because of its irrelevance.

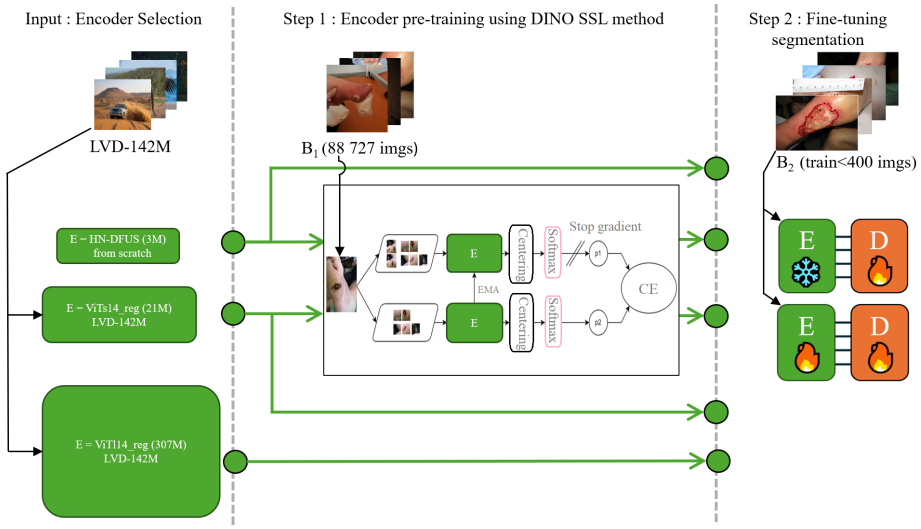

Fig. 2. Training scenario : the flake corresponds to the frozen encoder whereas flame corresponds to the use of the weight unfreezing strategy.

3.3 Implementation

The described experiments were conducted using an NVIDIA RTX A6000 graphics card with 336 Cuda Cores and 48 GB of memory.
SSL training with DINO is conducted using the library called lightly[5]. On-the-fly augmentation is performed during 300 epochs of SSL training with a mini-batch of 128 images, resulting in the creation of eight views. Resolution of the two "global" views is set to 224x224 and 98x98 for six "local" views. The projection head consists of three linear layers. Its input dimension depends on the output tensor dimension of each encoder, while the dimensions of the other layers remain unchanged between experiments, with values of 512, 128, and 2048, respectively. Regarding supervised training on B_2 coming after SSL pre-training on B_1, initial weights of encoders correspond to those that minimized the cross entropy loss function on SSL validation data. Segmentation is performed over 150 epochs

[5] https://github.com/lightly-ai/lightly

Table 1. DICE performances of models without SSL pre-training on B_2 test set supervised segmentation task. Results are given as mean ± std on 5 splits

Encoder	fine-tuning (B_2)	Train=70	Train=140	Train=280	line
HarDNet-DFUS$_{rdm}$	🔥	**0.69±0.04**	**0.74±0.03**	**0.77±0.02**	1
ViTs14_reg	❄	0.57±0.04	0.65±0.03	0.65±0.02	2
	🔥	**0.69±0.03**	0.72±0.02	0.73±0.01	3
ViTl14_reg	❄	0.64±0.04	0.64±0.02	0.70±0.03	4

Table 2. DICE performances of models with SSL pre-training on B_2 test set supervised segmentation task. Results are given as mean ± std on 5 splits

Encoder	fine-tuning (B_2)	Train=70	Train=140	Train=280	line
HarDNet-DFUS$_{rdm}$	❄	0.72±0.03	0.73±0.01	0.76±0.02	5
	🔥	**0.76±0.03**	**0.79±0.01**	**0.80±0.01**	6
ViTs14_reg	❄	0.60±0.06	0.64±0.04	0.67±0.03	7
	🔥	0.67±0.02	0.71±0.02	0.72±0.03	8

on five splits with the Lawin decoder: four feature maps are extracted from the encoder and adapted to the four expected inputs of the decoder.
Predictions are penalized by a loss function that combines weighted binary cross-entropy, weighted intersection over union, and also a boundary loss. This loss function uses the ground truth G, the final prediction P, intermediate predictions P_i at inner block i, and boundary prediction P_B. G_B corresponds to boundary ground truth. More details can be found in [9].

$$L = l^w_{BCE}(G,P) + l^w_{IoU}(G,P) + l_{BCE}(G_B,P_B) + \sum_{i=1}^{n}(l^w_{BCE}(G,P_i) + l^w_{IoU}(G,P_i))$$

The batch size depends on the GPU memory occupied by the encoder: 12 for HarDNet-DFUS and for ViT encoders when their weights are frozen, but 2 for ViT**s**14_reg during progressive unfreezing. Performance on the test set will be evaluated using the Dice metric, a commonly used measure in segmentation to assess the similarity between the model's prediction and the ground truth.

3.4 Results

During SSL training, an overfitting phenomenon occurs after about a hundred epochs, regardless of the encoder used. An early stopping strategy is set to 30 epochs to limit computation time, which in total corresponds to approximately 30 hours for one hundred epochs.
Tables 1 introduces the Dice performance on B_2 test set according to training scenarios that do not apply SSL pre-training on B_1. Conversely, table 2 refers to scenarios that do apply SSL pre-training on B_1. Figure 3 illustrates the same information but gathered according to fine-tuning choice during supervised segmentation on B_2. The left-hand side graph concerns scenarios that unfreeze encoders whereas the right-hand side graph illustrates scenarios that kept encoders freezed. Figure 4 illustrates the effect on wound segmentation of SSL pre-training on B_1 by showing the superposition of predicted contouring on the original picture.

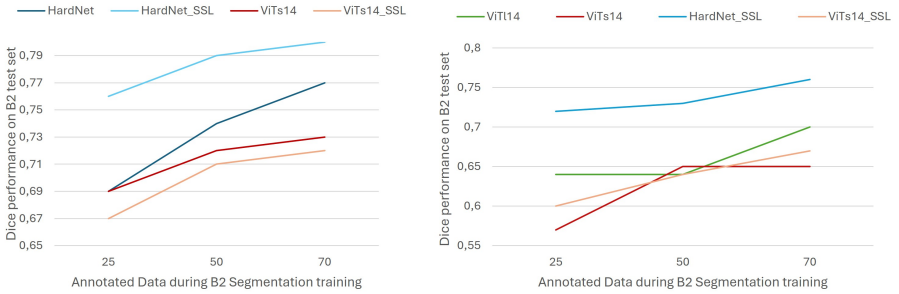

Fig. 3. Final Dice performances for scenarios in which encoders are unfreezed (left) or kept freezed (right) during supervised segmentation training on B_2.

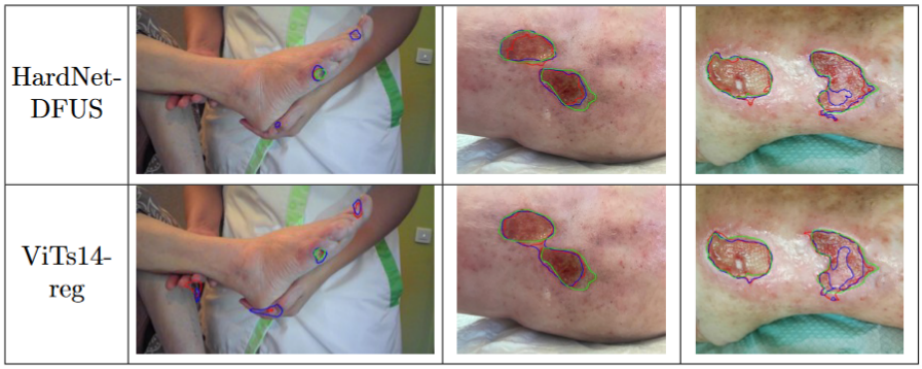

Fig. 4. Visualization of 3 inferences on test set with HarDNet-DFUS and ViTs14_reg. Models were fine-tuned for segmentation with B_2 version containing 280 training images. Blue is model without encoder pre-training on B_1. Red is model with encoder pre-training on B_1. Green is groundtruth. Thus, vizualisation corresponds to lines 1, 3, 6 and 8 from Table 1 and 2.

3.5 Discussion

When comparing all scenarios at once, HarDNet-DFUS encoder randomly initialized then pre-train on B_1 with the DINO SSL methodology, before supervised training on B_2 with unfreezing fine-tuning strategy, gives the best performances. Results presented in Table 1 demonstrate that increasing the scale of SSL DINO pre-trained ViT model leads to improved Dice metric. Indeed, scaling up the encoder from ViT**s**14_reg to ViT**l**14_reg (rising from 21 M to 307 M parameters) is associated with enhanced feature extraction capabilities for the encoder when it deals with chronic wound segmentation. One should also remark that a lightweight convolutional encoder such as the one from HarDNet-DFUS (3 M parameters), without undergoing pre-training, achieves better performance than ViT**l**14_reg initialized via SSL DINO methodology on LVD-142M. This observation is made regardless of the amount of training data during the final

supervised task. This suggests that LVD-142M is not suitable for clinical applications as specific such as chronic wound analysis.
Thanks to Table 2, we see that for the HarDNet-DFUS model, whatever the fine-tuning strategy during the supervised task we choose, applying the SSL DINO methodology on B_1 for pre-training results in a Dice metric improvement for the downstream segmentation task. Interestingly, ViT**l**14_reg performances are lower when a pre-training is done on dataset B_1 (around 88,000 pictures which is far smaller than LVD-142M generic database). This phenomenon is indeed classical and probably due to an insufficient number of training examples.
Figure 4 qualitatively shows the impact of SSL pre-training on B_1 for HarDNet-DFUS and ViTs14-reg models. First, we see that pictures are quite heterogeneous because they illustrate different types of chronic wounds with different acquisition conditions. With the left-hand picture, we can see SSL enables both models to limitate false positive segmentation. On the middle one, SSL enables models to differentiate the two wound instances. With the right-hand picture, we see SSL globally improve segmentation by making borders more accurate. In conclusion, using the SSL methodology on a dedicated database (i.e. B_1) with a small encoder (HarDNet-DFUS) is more interesting than using a huge one (ViT**l**14_reg or ViT**s**14_reg) pre-trained on a generic dataset (LVD-142M) for the specific segmentation task. Thus, using our proposition (pre-train with DINO methodology a small architecture on the application dataset before unfreezing the encoder during the downstream task) can lead to better performances. By the way, it also leads to annotation savings. Indeed, when the encoder is fine-tuned on B_2 with only 70 annotated images, HarDNet-DFUS model achieves similar performances (76% in Dice metric) compared to the same encoder but without pre-training on B_1, despite the use of 280 annotated data for its segmentation training.

4 Conclusion

In this paper, we have assessed the usefulness of using a DINO-trained encoder on a specific clinical dataset and segmentation task. We compared, in various training scenarios, the performances of two encoders: the ViT encoder, whose weights are derived from the Dinov2 paper, to the lightweight one from the HarDNet-DFUS model, initialized with random weights. Results indicate that it is not necessary to use generically pre-trained DINO ViT encoders since conventional lightweight ones may outperform them on specific tasks. Additionally, pre-training an encoder via DINO SSL on a specific dataset with limited annotations proves to be beneficial for small encoders. It would be interesting to further explore this topic by investigating the impact of increasing the database on SSL training through the addition of all public databases related to dermatological lesions.

Acknowledgments. We would like to thank the National Association for Research and Technology as well as Cicat-Occitanie for supporting the CIFRE thesis.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[2] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924.

[3] Bill Cassidy et al. "The DFUC 2020 Dataset: Analysis Towards Diabetic Foot Ulcer Detection". In: *European Endocrinology* 1.1 (2021), p. 5. ISSN: 1758-3772. DOI: 10.17925/ee.2021.1.1.5. URL: http://dx.doi.org/10.17925/EE.2021.1.1.5.

[4] Ping Chao et al. "Hardnet: A low memory traffic network". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3552–3561.

[5] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv:2010.11929* (2020).

[6] Chien-Hsiang Huang et al. *HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS*. 2021. arXiv: 2101.07172 [cs.CV].

[7] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[8] Connah Kendrick et al. "Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation". In: *arXiv:2204.11618* (2022).

[9] Ting-Yu Liao et al. "HarDNet-DFUS: Enhancing Backbone and Decoder of HarDNet-MSEG for Diabetic Foot Ulcer Image Segmentation". In: *Diabetic Foot Ulcers Grand Challenge*. Springer, 2022, pp. 21–30.

[10] Laura Martinengo et al. "Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies". In: *Annals of epidemiology* 29 (2019), pp. 8–15.

[11] Maxime Oquab et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv:2304.07193* (2023).

[12] Utku Ozbulak et al. "Know Your Self-supervised Learning: A Survey on Image-based Generative and Discriminative Training". In: *arXiv:2305.13689* (2023).

[13] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[14] Zhe Wu et al. *Cascaded Partial Decoder for Fast and Accurate Salient Object Detection*. 2019. arXiv: 1904.08739 [cs.CV].

[15] Haotian Yan et al. "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention". In: *arXiv:2201.01615* (2022).

[16] Chuyan Zhang et al. "Dive into the details of self-supervised learning for medical image analysis". In: *Medical Image Analysis* 89 (2023), p. 102879.

[17] Xizhou Zhu et al. *Deformable ConvNets v2: More Deformable, Better Results*. 2018. arXiv: 1811.11168 [cs.CV].