Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(9), 163–174
https://doi.org/10.12913/22998624/224196
ISSN 2299-8624, License CC-BY 4.0

Received: 2026.05.23
Accepted: 2026.07.10
Published: 2026.08.01

Superpixel-graph self-supervised pretraining with joint-embedding and contrastive objectives for wound image segmentation

Bohdan Lukashchuk[1*], Ihor Farmaha[2]

[1] Computer Science Department, Ukrainian National Forestry University, 103 Gen. Chuprynky St., Room 45, 79057 Lviv, Ukraine
[2] Department of Computer Design Systems, Lviv Polytechnic National University, 3 Metropolitan Andrey St., Building IV, Room 324, 79013 Lviv, Ukraine
* Corresponding author's e-mail: bohdan.lukashchuk@gmail.com

ABSTRACT

Annotated data remain a major problem and bottleneck for supervised learning approaches, especially in wound imaging. In this work, a self-supervised learning pretraining method was proposed that represents a wound image as a superpixel adjacency graph, generated using the simple linear iterative clustering algorithm and uses graph structure to define building blocks for calculating joint-embedding and contrastive learning objectives. Three objective variants were evaluated – contrastive, joint-embedding and their combination as a pretraining stage for further finetuning for the wound segmentation task. After self-supervised pretraining on 100 unlabelled images and supervised fine-tuning on only 30 labelled images using the contrastive objective, a Dice score of 0.266 was received on a 400-image test set, compared to 0.354 for a fully supervised U-Net trained on all 100 annotated images, thus having a gap of 0.088 in Dice score with more than three times less annotated data. The results suggest that superpixel-graph-based self-supervised learning is a promising pretraining strategy for wound analysis in settings with limited annotated data.

Keywords: self-supervised learning, joint-embedding predictive architecture, superpixel segmentation, deep learning, convolutional neural networks.

INTRODUCTION

Computer vision (CV) has been widely used for different task automation in the healthcare sector for some time now. [1]. Typical use-cases include the utilisation of CV methods to analyse CT, MRI and PET data [2]. In recent years, more researchers and practitioners have explored the use of computer vision for the tasks of wound imaging [3, 4]. Open-source datasets of wound images [5] and tools to ease and automate dataset annotation [6] are being created. This is motivated by some very practical needs:

- wound classification, which is required for faster clinical decision making. Example – wound classification into one of: diabetic, pressure, surgical, venous ulcers [3, 4];
- wound segmentation (wound boundary delineation) – helps in monitoring the healing dynamics. Example – ulcer wound area segmentation [5, 7];
- wound tissue classification – next step further from wound classification. Example - classification of wound tissues: epithelialisation, granulation and necrotic tissues [8].

The majority of the aforementioned tasks are being approached using deep learning CV models (variants of convolutional neural networks [9] (CNN) or vision transformers [10] (ViT)) that rely on the supervised learning approach [11]. For supervised learning, annotated data is crucial

as the whole strategy relies on a "tuning model parameters in a way that minimises error between predicted and ground truth values" principle. This kind of learning requires large amounts of annotated data. In a wound imaging setting, this means collecting hundreds (better thousands) of images, which may require obtaining different legal permissions from the patients, hospitals, etc. and then annotating them – either allocating class labels for each of the images or additionally delineating areas of interest (wound area, specific tissue area, etc.), which can be correctly done only by the doctors and require lots of time and financial resources. Another common issue with supervised learning methods is a lack of generalisation to new data. In one of authors' previous research projects, a concrete failure case was reported: a supervised segmentation model (U-Net), trained on the FUSeg dataset for the wound image segmentation task, can reach a high Dice score on an in-domain foot ulcer dataset while being ineffective on a different wound dataset collected under different conditions [12].

The aforementioned problems with data and models' poor generalisation abilities, when trained in a traditional supervised manner, motivated the authors to explore unsupervised, self-supervised, and few-shot learning methods that can be adapted to wound analysis tasks.

This manuscript is a continuation of two of authors' previous studies. First [13], it was investigated that the unsupervised SLIC [14] method can generate reliable superpixels for wound image segmentation, i.e. able to determine correct SLIC hyperparameters to generate superpixels that, when combined, have boundaries that coincide with the wound boundaries. Second [12] – unsupervised (autoencoder-based) and few-shot (Siamese/contrastive learning) feature learning from random image crops were explored, aiming to obtain the representations that clearly separate wound, skin, and background while generalising well with limited annotated data. The core outcome of the latter research is that contrastive training produced the most discriminative embeddings (patch classification accuracy above 93% across variants). At the same time, variational and classical autoencoders were unstable for discrimination in this setting. However, the first study outcome still leaves the problem of superpixel classification unresolved, which is traditionally a supervised learning task that requires annotated data, and the second study outcome shows that using unsupervised (reconstructive) approaches with autoencoders is insufficient. Contrastive learning, while producing promising classifications with only a few-shot learning, may fail to generalise to unseen data. These results motivated us for the current project. Here, the authors delved deeper into the self-supervised learning techniques and proposed a superpixel-aware self-supervised pretraining strategy that treats the wound image as a graph of superpixels and learns separate superpixel representations by employing a joint embedding approach from JEPA [15] and a discriminative approach of SimCLR (contrastive learning) [16] thus both learning to distinguish the superpixel representations of the known domain (wound images) and representations "inside" wound images (e.g. wound, skin, background, etc.).

While some annotated training data is available, it is limited in both quality and quantity, and that effective self-supervised pre-training is needed to enable the subsequent few-shot supervised fine-tuning for the task of interest.

Recent years have shown the capabilities of large language models (LLMs). Arguably, the key to their success is how they are trained. GPT-style [17] decoder models – given sequence of tokens $(x_1, ..., x_{t-1})$ predict next token x_t. BERT-style [18] encoder models use masked language modelling: mask a few tokens from the sequence and predict them from the masked input sequence (very simplified explanation). Basically, the idea of these methods is to derive labels from the data – hide part of the data and make the model learn how to predict the hidden target data from the context – everything else. This makes the model domain aware – learn how the data is distributed across the specific domain. In the case of LLMs, this worked very well, as raw text data is available all over the Internet. This and even older approaches from natural language processing, like Word2Vec's skip-gram [19]. The model, combined with ideas from contrastive and metric learning, inspired computer vision researchers to develop self-supervised learning methods.

The most prominent self-supervised learning methods in vision can be divided into two main groups:

1. Contrastive learning.

They are built around the idea of discriminating instances that are considered similar

("positive") to the "anchor" view versus all or some portion of the other views – "negatives". The authors investigated the topic and Nt-Xent loss [16] in their previous work [12]. The most notable contrastive methods:

- SimCLR [16], where two "positive" views of the "anchor" image are created using augmentations of the "anchor" image and "negative" views are all the other images in the training batch. The NT-Xent loss is used to push the embeddings of the "positive" views closer to the "anchor" view and further from the "negative" views. This method plays a great role in our research.
- MoCo [20] – query and key – two augmented views of the same image. During training, the query is pushed closer to its *key* and far from the keys of the other images. However, the difference is that a special queue stores keys from many previous images (not only the current batch). Uses two encoders: one, trained in a typical way, produces query embeddings, and the second is a copy of the first but is updated much more slowly and generates key embeddings. This trick is used to ensure *key* consistency during training. Many other self-supervised approaches use separate networks for different embeddings.

2. Joint-embedding learning.

Unlike the previous group, this focuses on bringing similar image embeddings closer together without using negatives, so nothing is pushed too far. Collapse is prevented using different architectural or regularisation techniques.

- BYOL [21] – very similar to the MoCo, two augmentations of the same image, two encoders, one fast (online network), another slow, learning not via gradient descent but copying the exponential moving average (EMA) of the weights of the online (target network). One view is passed over the network and then through the predictor head; the other is passed through the target. Embeddings are pushed closer, but no negatives are used. The approach is very similar to MoCo, but without negatives.
- DINO [22] – few crops are made from the same image: 2 large crops (global views) and many small crops (local views). Two networks: the student is updated via backpropagation, and the teacher uses a copy of the student's EMA weights (a common approach in self-supervised learning). Global views are passed through the teacher network, which produces probability over classes (soft classification). Global and local views are then passed through the student, which also outputs probability vectors. Objective: make the student model output equal to the teacher vector for the same network. This one has an important extension from the methods mentioned before – it works not with the full images and their augmentations, but with the different levels of image crops, thus enhancing local feature detection.
- I-JEPA [15] – split image into patches via a grid, mask large chunks of patches. Two encoders – context and target encoder, the latter uses EMA weights of the former. Pass the visible patches through the context encoder to obtain the context embedding, then pass the embedding through the predictor to refine it. Pass the entire image through the target encoder to obtain the target embedding. Loss makes the predictor (context) embeddings match the target embeddings. Conceptually, the approach is similar to masked modelling in BERT-like LLMs. This is the second piece of research that significantly influences this project.

Some methods use generative or clustering approaches, but the two aforementioned groups are the most influential. As shown in the method list above, many self-supervised methods in vision use two-network training, where one network is trained by transferring EMA weights from the other. This is not the only hack to ensure training stability and avoid embedding collapse in multidimensional space when training on the large general-purpose datasets. Some studies, like LeJEPA [23], try to eliminate this problem mathematically by forcing embeddings to an isotropic Gaussian distribution.

Self-supervised learning approaches in the wound imaging domain are underrepresented in the research literature. The authors of [24] use self-supervised encoder pretraining for chronic wound segmentation, using their own private dataset of 133000 images with very few pixel-level annotated masks. The authors selected 88727 images for SSL pretraining and 400 annotated images for supervised training/fine-tuning. They compared a ViT pretrained with the DINOv2 [25] approach on a large general-purpose dataset, their custom small HarDNet-DFUS encoder trained from scratch on their

annotated dataset, and the same HarDNet-DFUS encoder after in-domain DINOv2 SSL on their unlabelled wound dataset. They followed pretraining with the supervised fine-tuning on very few masks. The main outcome of the study was that in-domain SSL pretraining yielded the largest gains for the small custom model, greatly reducing the need for annotation masks (0.76 Dice score with 70 masks available, comparable to completely supervised training with 280 annotated images). This is an important study, as it demonstrates the capabilities of using small custom models with SSL approaches in the wound imaging domain. The authors of [26] use a different SSL approach: they train the model on 110 expertly pixel-level annotated images of foot ulcer tissues (granulation, fibrin, callus) in a supervised manner, and also use 600 unlabelled images. Trained on the labelled dataset, the model generates pseudo-labels (predictions) for some of the unlabelled examples and uses these pseudo-labelled annotations in subsequent training iterations. If the validation loss increases, more pseudo-labelled images are added to the training set. Authors used a special combination of Dice loss and Focal loss [27]. They achieved a Dice Score of 87.64% with their pseudo-label SSL pipeline, compared with 84.89% with supervised-only training. In [28], the authors do not train models in an SSL manner but utilise already pretrained by DINOv2 models and fine-tune them for wound-type classifications. It is also worth noting that SSL utilisation is used in medical imaging but not in wound imaging. Authors of [29] utilise a masked image modelling SSL approach in the histopathology domain. In [30], contrastive learning (MoCo v2 [31]) is used to pretrain models on chest X-ray images, followed by fine-tuning on the pleural effusion classification task. However, instead of using two separate image augmentations as positive pairs, as in MoCo, they use two images that are very likely to have the same pathology. The adjusted MoCo approach is also used by the authors of [32] for X-ray imaging – pleural effusion detection. A slightly different approach was used by the authors of [33]. They utilised a contrastive approach for the X-ray imaging, but instead of two augmented views of the same image, they combined X-ray image embedding with its radiology report (text) embedding, this is an example of vision language SSL modelling in the medical domain. Overall, these studies show that in many domains of medical imaging, researchers try to utilise pretraining on the unlabelled data using SSL techniques to eliminate the need for expensive expert annotations.

The goal of this work was to develop and evaluate a superpixel-graph-based self-supervised pretraining method for learning wound image representations from unlabelled data, and to assess whether the pretrained encoder can improve subsequent wound segmentation under limited labelled data.

MATERIALS AND METHODS

Data

There are a few publicly available datasets for wound image analysis:

- FUSeg dataset of 1210 foot ulcer images and corresponding wound segmentation masks [5].
- WoundSeg dataset of 2686 wound images and corresponding segmentation masks [34].
- Medetec dataset of 595 wound images, divided into a few wound classes (no segmentation masks) [35].
- QTDU dataset of 44893 superpixels, annotated to one of the four classes: wound tissue: fibrin (3974), granulation (3284), necrotic tissue (448), healthy skin (37187) from 40 images of arterial and venous wounds from different and unrelated patients [36].

While at first glance the amount of data looks solid, close analysis of the datasets reveals multiple problems in the data:

- WoundSeg dataset has incorrect annotations [37] as well as discrepancies with licenses, where in one place [38] MIT license is stated, which allows any use and in the other dataset location [37] Creative Commons Attribution-NonCommercial 4.0 International (cc-by-nc 4.0) is used, which prohibits commercial use
- Medetec has a large class imbalance and no segmentation masks.
- QTDU has a large class imbalance and superpixels of very low quality, gathered only from 40 different wound images.
- FUSeg does not provide info on license (assuming free to use); however, it contains the data only on foot ulcer wounds.

The proposed method combines SSL ideas of joint-embedding predictive modelling with

contrastive learning. It utilises superpixels – semantic patches that form the boundaries of different objects on the image, to create target and context as well as positive and negative views for the proposed SSL pretraining method.

Wound image as a superpixel graph

Let us define an RGB image as: $I \in \mathbb{R}^{3\times H\times W}$ where H and W are the image height and width, respectively, for each image in the training set, we compute hard segmentations using the SLIC method:

$$S = \{S_n\}_{n=1}^{N_{sp}} \tag{1}$$

where: the number of superpixels N_{sp} = max(H, W), as per authors' previous research [13].

Now, let us define the superpixel adjacency graph. $G = (V, E)$, where each node $v_n \in V$ corresponds to superpixel S_n and edges connect superpixels that share common boundaries (Figure 1).

Node representation

Each node v_n is an embedding vector, generated from the output features of authors' custom convolutional encoder. The image is passed through a lightweight U-Net encoder [39] f_θ with three downsampling stages L that map image I to a feature map Equation 2.

$$F = f_\theta(I) \epsilon \mathbb{R}^{C\times H'\times W'} \tag{2}$$

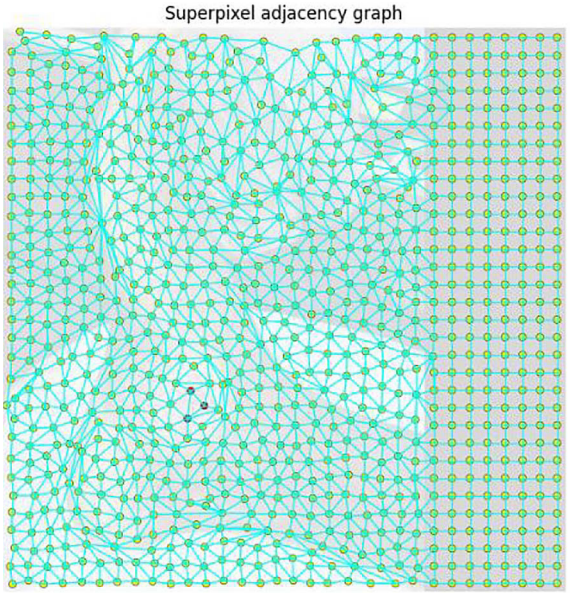

Figure 1. Superpixel adjacency graph formed from superpixels, generated on the foot ulcer wound image

where: H' and W' are the corresponding height and width of the resulting feature map and $H' = H/2^L$, $W' = W/2^L$, $L=3$, C is the number of channels at the output feature map, and $C = 8 \times 2^L$ (Figure 2).

A 512 × 512 × 3 input image I via encoder is transformed to the 64 × 64 × 64 feature map F. After that, the feature map is upsampled to 512 × 512 and, for each superpixel, the 64-dimensional feature vectors are average over all its pixels:

$$v_i = \frac{1}{S_i}\sum_{(y,x)\in S_i} F \uparrow (:, y, x), \tag{3}$$

By stacking the 64-dimensional vectors received, a complete vector representation of the graph nodes is received:

$$V = \begin{bmatrix} v_1 \\ v_2 \\ . \\ . \\ . \\ v_{N_{sp}} \end{bmatrix} \in \mathbb{R}^{N_{sp}\times C} \tag{4}$$

Both the individual feature embeddings from the feature map F and the adjacency graph G play important roles in the subsequent steps.

Joint-embedding objective

I-JEPA-like complex training protocol with a two-encoder setup and weight transfer from one network to the other was not employed, but their idea of training objective was used, which, in general, is as follows:

- define large context view s_c (for example, a large random crop from the image I that covers almost the whole image I);
- define small target view s_t (for example, a small crop of image I);
- *context* and target views do not overlap $s_c \cap s_t = \varnothing$;
- pass each of them through student and teacher encoder networks and receive corresponding embeddings: $z_c = f_q'(s_c)$, $z_t = f_\theta(s_t)$;
- pass context embedding through a small predictor – multi-layer perceptron (MLP): $z_c' = g_\theta(z_c)$;
- enforce z_c' and z_t similarity with the L2 loss.

The idea of *target* and context embeddings was taken and combined with the superpixel adjacency graph G. During training, after passing the whole image I through the encoder, f_θ (use only

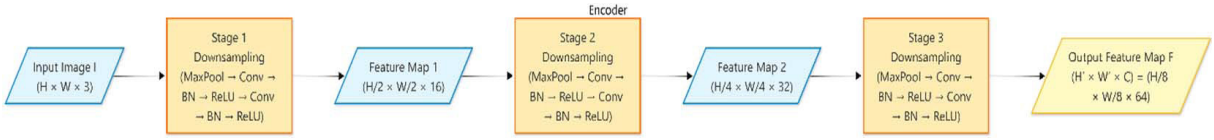

Figure 2. Encoder architecture

one encoder for simplicity), randomly (with equal probability for each) one of the two possible set-ups for target and context views was selected:

1. Target s_t is some (randomly selected) superpixel (node v_n) and its 1-ringh neighbours N(v_n). *Context* s_c – all other superpixels of the image, except for the target ones. This forces the model to learn global features from local ones.
2. Target s_t is some (randomly selected) superpixel (node v_n) and *context* s_c - its 1-ringh neighbours N(v_n). This enforces local smoothness.

After this, the corresponding z_c, z_t embeddings were received using average pooling, as described previously, and z_c' was received, after passing z_c through the predictor MLP and optimising *L2* loss between z_t and z_c':

$$L_{joint} = (s_t - s_{c'})^2 \tag{5}$$

The authors opted with the simpler one-encoder approach and not use complex training methods, as their main purpose in I-JEPA is to ensure stability during training on very large image datasets from the whole Internet, and for training on smaller datasets, a single-domain one-encoder is sufficient.

The authors suggest that, because superpixels already represent small but semantic regions of the image and can outline objects individually or in combination with adjacent superpixels, defining *target* and *context* views via superpixels, rather than random rectangular patches, eases network training for general understanding.

Contrastive objective

While the joint-embedding objective provides in-domain learning, authors' initial experiments show that it is not sufficient for good object discrimination. Thus, the authors decided to use the experience from their previous studies [12] and integrate a contrastive learning regularisation term into the joint-embedding objective. Contrastive learning requires the definition of anchor, positive and negative views. In SimCLR, the anchor is a full image itself; positives are its augmentations, and negatives are all other images in the minibatch. The work is not carried out at the image level but at the superpixel level, thus defining all these in terms of the adjacency graph *G* nodes. The superpixel node v_c embedding was used, selected on the previous step (target and context definition) as an anchor and its 1-ring neighbours as positives. Then negatives were selected (same amount as positives or less, if not enough nodes available), uniformly distributed from all the other nodes in the graph, which are further than a specific graph distance (configurable parameter d) from the anchor node and calculate NT-Xent loss on all of these embeddings:

$$L_{contr} = -\frac{1}{|\mathcal{P}|}\sum_{j\in\mathcal{P}} log\frac{\exp\left(\frac{sim(v_c, v_j)}{\tau}\right)}{\sum_{k\in\mathcal{P}\cup\mathcal{N}}\exp\left(\frac{sim(v_c, v_k)}{\tau}\right)} \tag{6}$$

where: $\mathcal{P}(i)$ – set of positives for anchor v_c (1-ring neighbours), $\mathcal{N}(i)$ – set of sampled negatives for anchor (nodes that are far away), v_j – positive sample, v_k – either a positive or a negative sample, τ – temperature parameter. For consistency with the joint-embedding loss as a measure of distance sim, the L2 loss was used.

Figure 3 shows the visual representation of the selection method. The red area around the anchor and positives exactly covers the nodes that are ≤ the predefined graph distance *d*. However, two questions arise when considering this approach. How to prove that 1-ring neighbours should be selected as positives, and how to choose the optimal d – "non-negative" zone radius around the anchor so it maximises the correct separation for the current domain, so that wound superpixels are more likely to be grouped with the other wound superpixels? To tackle this problem, a small calibration was run on 100 annotated wound images.

Superpixels were labelled as wound if at least 80% of their pixels covered the wound area in the image. For each image, a superpixel adjacency graph was built and randomly

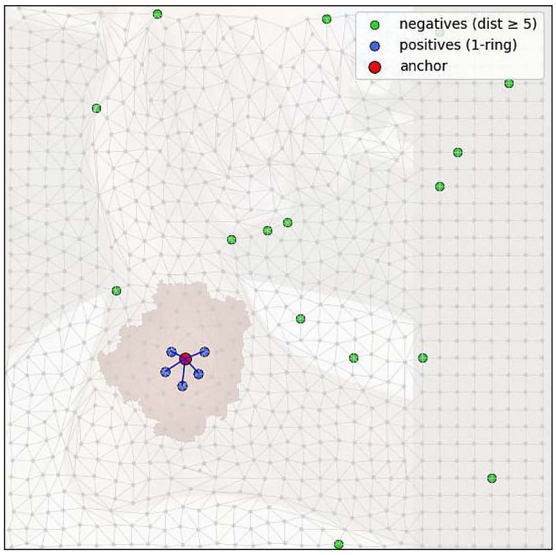

Figure 3. Visualisation of anchor, positives and negatives selection for the contrastive regularisation term

selected 200 anchors for the test. After that, the authors computed:

- 1-ring label purity – the fraction of 1-ring neighbours that share the anchor's label;
- boundary edge rate – the fraction of edges connecting nodes of different labels (global and for edges incident to wound nodes).

The results show near-perfect local consistency for all anchors (global 1-ring purity 0.987, median 1.0), confirming that 1-ring neighbours are a reliable source of positive pairs. However, it should be noted that this value is largely due to the fact that wounds occupy a very small portion of the image area, resulting in a severe class imbalance that affects the 1-ring label purity. For wound anchors, purity is lower (0.809) and depends on wound size (by the amount of wound superpixels on the image): 0.738 for small wounds (20–49 superpixels), 0.866 for medium (50–149), and 0.918 for large (≥150), indicating that deviations mostly occur near wound boundaries in small wounds, which is logical, as small wounds may consist only of the boundary superpixels. However, even 0.738 purity for the small wounds is strong evidence that the 1-ring neighbour is a stable selection for positives.

Globally, boundary edges are rare (0.0129 of all edges), which may also be caused by the class imbalance (wounds are not large). In contrast, for wound-incident edges, the boundary-crossing fraction is much higher (0.300, per-image mean 0.575), which is consistent with many wound superpixels lying on the wound contour and with the statement that wounds take little space on the image.

To choose d, measuring label consistency was proposed across different d-rings: the proportion of superpixels that are d-ring neighbours of the anchor superpixel and share the same label as the anchor. The authors tested $d \in \{3, 5, 7\}$ and obtained 0.547, 0.372, and 0.263, respectively, so 3 appears to be the optimal value based on the current test. However, empirical tests show that using d=5 in many cases yields better results; thus, the d-selection approach needs further improvement.

Combined objective and experiment setup

The combination of contrastive and joint losses is as follows:

$$L_{total} = L_{jepa} + \lambda L_{contr} \tag{7}$$

where: λ is a multiplier for the contrastive objective and, based on the empirical evidence, is set to 0.1. Temperature parameter τ for contrastive loss, it is set to 0.2 as per [12].

As primary data source, a joint wound dataset was used [40], which is a combination of the aforementioned FUSeg, WoundSeg and Medetec datasets and contains 2760 wound images and annotation masks (Medetec annotations are done by the author of [40]).

To evaluate all three objectives (JEPA, contrastive and their combination), the authors built on top of the superpixel graph and develop a controlled and reproducible protocol:

- SSL pretraining – for each of the three seeds $s \in \{0,1,2\}$ select 100 non-overlapping images from [40] train set;
- for each of the objectives (jepa, contrastive, combination) run SSL pretraining (learning rate 1e4, 20 epochs);
- for each objective, select the best (lowest loss) model across the seeds;
- define test set – 400 wound images with annotations from [40] test set and extract superpixels from it;
- for each of the best models visualise activations on the random images from the test set;
- visualise final model layer embeddings (of superpixels from 100 random images from the test set) using PCA for dimensionality reduction.

After visual analysis of the SSL pretraining step, a U-Net-like decoder with randomly initialised weights was added and tested how different pretrained encoders affect supervised fine-tuning for the semantic segmentation task. Weighted binary cross-entropy and Dice loss were used:

$$L_{sup} = \lambda L_{dice} + (1 - \lambda) L_{bce} \quad (8)$$

where: L_{dice} – Dice loss, L_{bce} – binary cross-entropy loss and λ-0.5 – weighting multiplier.

- supervised finetuning for each of the seeds s – define 4 supervised finetuning datasets, each with K images, $K \in \{1,3,5,30\}$, from the 100-image seed dataset;
- for each of the supervised datasets – finetune model in a supervised manner on top of the current JEPA, contrastive and combined SSL models (learning rate 1e4, 20 epochs);
- evaluate models using Dice and IoU metrics;
- for each seed s, treat a 100-image dataset as a supervised dataset, split it into training and validation sets at an 80/20 ratio, and train a classical supervised segmentation model (U-Net). Then compare the IoU and Dice scores with those of the fine-tuned models.

RESULTS

The authors decided to use 100-image dataset for the SSL finetuning because current implementation of the target, context, positive and negative views selection based on superpixel adjacency graph allows only 1-batch training and pretraining process, while not being resource-intensive in terms of computational capabilities, takes a lot of time – approximately 3 hours on Intel Core i7-9750H CPU 22.59 GHz, 32 GB DDR4 RAM, Nvidia 2060 6 GB VRAM per one training session (20 epochs, 1 loss).

Figure 4 shows activations from the different stages of the pretrained SSL encoder, using contrastive, JEPA, and combined losses, respectively.

The presented visualisations of the activation functions do not indicate strong segmentation signals across any of the proposed objectives. However, training with a combined objective on the 100 epochs (compared to the 20), shows better activations, which imply good segmentation ability on some of the images (Figure 5). Figure 6 shows superpixel embeddings in a scatter plot, with dimensions reduced from 64 to 2.

Contrastive and combined objectives show some signals of wound representation grouping.

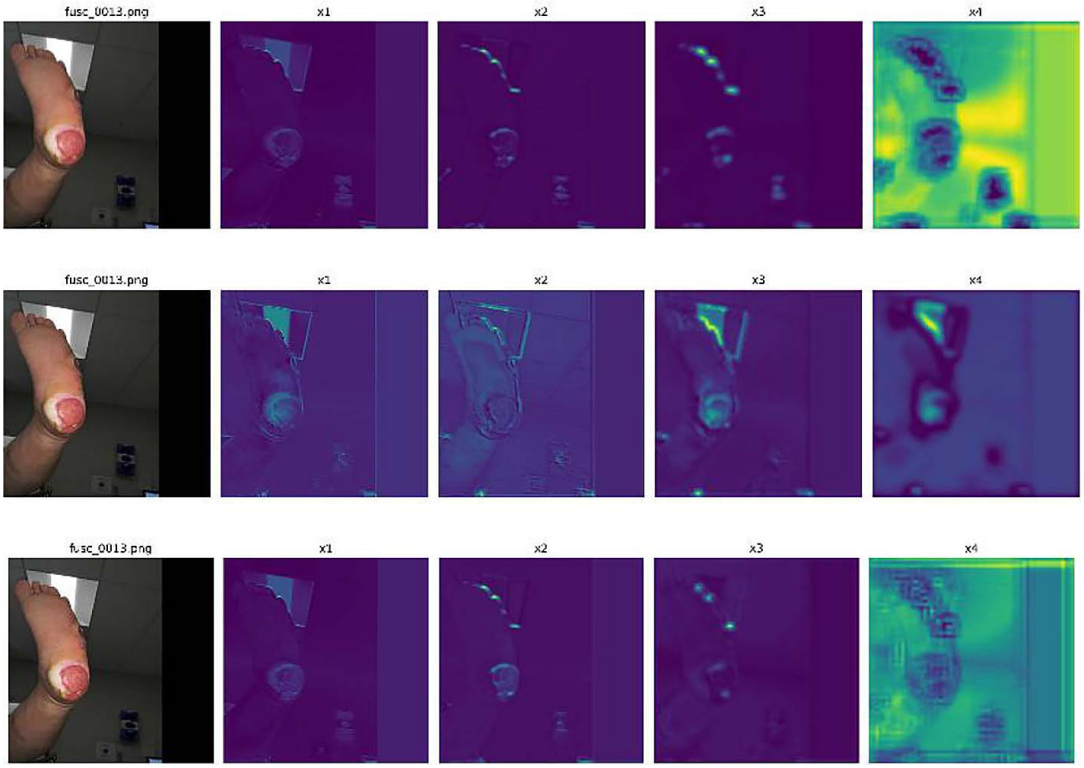

Figure 4. Visualisation of activation maps, top to bottom: contrastive, JEPA, combined JEPA and contrastive SSL encoder after 20 epochs of training

Still, they are much more scattered than the JEPA embeddings, which, in authors' opinion, is caused by the repulsive part of the contrastive loss. However, they are still completely overlapped with the other superpixels. Thus, it cannot be concluded that SSL pretraining alone, in any of the three variants, can achieve good wound vs all other superpixel separation.

Supervised finetuning Dice metric results are presented in Table 1 and IoU results in Table 2 (averaged over three splits). In a nutshell, these results indicate that few-shot finetuning on top of the proposed SSL pretraining methods with a dataset size $K \in \{1,3,5\}$ do not provide sufficient generalisation capabilities on the 400-image test set, achieving a maximum metric value of 2.8%. Potentially promising results start to appear at K=30. Contrastive objective learning achieved a mean Dice score of 26.6%. For comparison, the supervised U-Net model, trained on 100 training images, achieved a Dice score of 35.4% and an IoU of 27.9%. Two of the metrics are approximately 10% better than the best fine-tuning results; however, the amount of supervised data used is 3 times larger than in the adopted SSL and fine-tuning settings.

When comparing SSL methods, supervised fine-tuning on top of contrastive SSL pretraining yields the best results. However, the ambiguity of the activation map and PCA visualisations suggests that longer pretraining sessions (200–300 epochs) are needed to obtain more empirical evidence.

CONCLUSIONS

In this manuscript, the authors presented their SSL pretraining method based on the superpixel adjacency graph and its utilisation for selecting

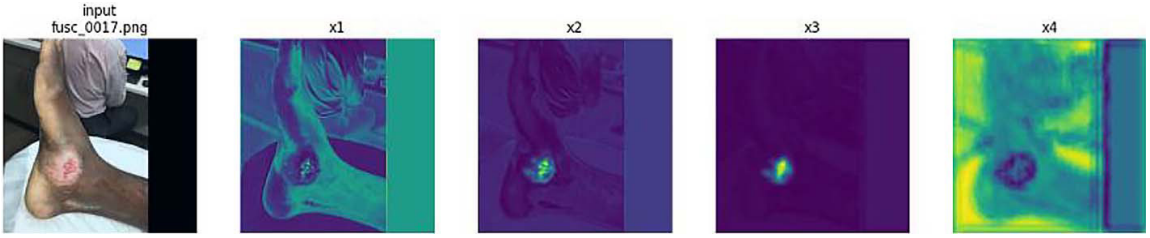

Figure 5. Visualisation of activation maps, combined JEPA and contrastive SSL encoder after 100 epochs of training

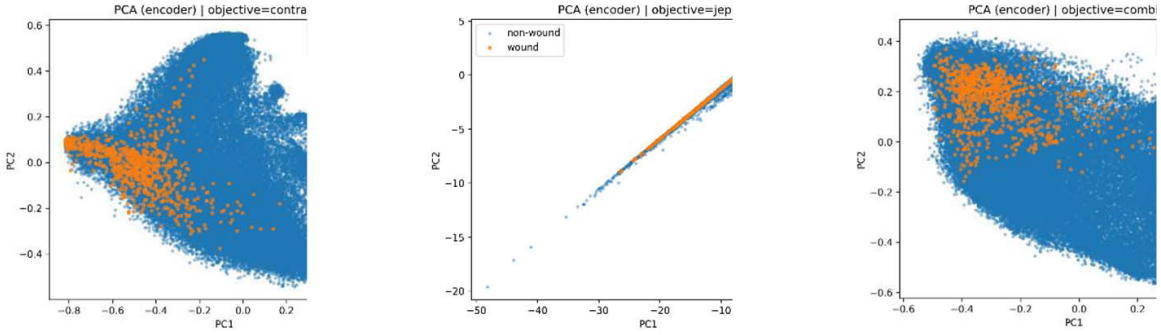

Figure 6. Visualisation of superpixel embeddings using PCA (orange – wound, blue – non-wound)

Table 1. Mean Dice score metric values for supervised finetuning on the datasets with different size K

K Objective	1	3	5	30
Contrastive	0.024	0.109	0.143	0.266
Jepa	0.028	0.087	0.080	0.183
Combined	0.062	0.089	0.170	0.193

Table 2. Mean IoU metric values for supervised finetuning on the datasets with different size *K*

K Objective	1	3	5	30
Contrastive	0.013	0.065	0.092	0.180
Jepa	0.015	0.052	0.046	0.121
Combined	0.038	0.053	0.108	0.129

target, context, positive, and negative views for JEPA-like and contrastive objectives, and validated it on the wound segmentation problem. The superpixel graph formulation provides a semantically meaningful alternative to random rectangular patches for view selection, as confirmed by authors' study, showing 0.987 global 1-ring label purity and 0.809 purity for wound-specific anchors. The authors experimented on JEPA-like and contrastive objectives combinations and in a best case received 0.266 Dice score (400 test images), after SSL (no label) pretraining on 100 images and further supervised finetuning on 30 of them, compared to 0.354 Dice score at fully supervised training on 100 images (with labels), thus having 0.088 difference in metric values and more than 3 times different in annotated data requirement. The main limitation of the current implementation is its long execution – the batch-size-one superpixel graph construction restricts training throughput, resulting in approximately 3 hours per 20-epoch session on a mid-range laptop GPU. The 100-epoch experiment with the combined objective produced noticeably improved activation maps, suggesting that longer SSL pretraining can enhance wound-relevant representation learning.

REFERENCES

1. Javaid M., Haleem A., Singh R.P., Ahmed M. Computer vision to enhance healthcare domain: an overview of features, implementation, and opportunities. Intell. Pharm. 2024; 2(6): 792–803. https://doi.org/10.1016/j.ipha.2024.05.007
2. Dayarathna S., Islam K.T., Uribe S., Yang G., Hayat M., Chen Z. Deep learning based synthesis of MRI, CT and PET: review and analysis. Med. Image Anal. 2024; 92: 103046. https://doi.org/10.1016/j.media.2023.103046
3. Reifs D., Casanova-Lozano L., Reig-Bolaño R., Grau-Carrion S. Clinical validation of computer vision and artificial intelligence algorithms for wound measurement and tissue classification in wound care. Inform. Med. Unlocked. 2023; 37: 101185. https://doi.org/10.1016/j.imu.2023.101185
4. Patel Y., Shah T., Dhar M.K., Zhang T., Niezgoda J., Gopalakrishnan S., et al. Integrated image and location analysis for wound classification: a deep learning approach. Sci. Rep. 2024; 14: 7043. https://doi.org/10.1038/s41598-024-56626-w
5. Wang C., Mahbod A., Ellinger I., Galdran A., Gopalakrishnan S., Niezgoda J., et al. FUSeg: the foot ulcer segmentation challenge. Information. 2024; 15(3): 140. https://doi.org/10.3390/info15030140
6. Jaworski N., Farmaha I., Marikutsa U., Farmaha T., Savchyn V. Implementation features of wounds visual comparison subsystem. In: 2018 XIV-th International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH); 2018; 114–117. https://doi.org/10.1109/MEMSTECH.2018.8365714
7. Farmaha I., Banaś M., Savchyn V., Lukashchuk B., Farmaha T. Wound image segmentation using clustering based algorithms. New Trends Prod. Eng. 2019; 2(1): 570–578. https://doi.org/10.2478/ntpe-2019-0062
8. Chairat S., Chaichulee S., Dissaneewate T., Wangkulangkul P., Kongpanichakul L. AI-assisted assessment of wound tissue with automatic color and measurement calibration on images taken with a smartphone. Healthcare. 2023; 11(2): 273. https://doi.org/10.3390/healthcare11020273
9. LeCun Y., Boser B., Denker J.S., Howard R.E., Hubbard W., Jackel L.D., Henderson D. Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems 2. San Francisco (CA): Morgan Kaufmann Publishers Inc.; 1990; 396–404.
10. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv [Preprint]. 2021 Jun 3. https://doi.org/10.48550/arXiv.2010.11929
11. Jiang T., Gradus J.L., Rosellini A.J. Supervised machine learning: a brief primer. Behav. Ther. 2020; 51(5): 675–687. https://doi.org/10.1016/j.beth.2020.05.002
12. Lukashchuk B. Computer modelling of textures on images with human skin wound areas. Int. J. Comput. 2024; 23(3): 486–497. https://doi.org/10.47839/ijc.23.3.3669
13. Lukashchuk B., Shabatura Y. Methodology for evaluating complex object contour detection

accuracy in SLIC-based image segmentation. Sci. Bull. UNFU. 2024; 34(8): 109–119. https://doi.org/10.36930/40340813

14. Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 2012; 34(11): 2274–2282. https://doi.org/10.1109/TPAMI.2012.120
15. Assran M., Duval Q., Misra I., Bojanowski P., Vincent P., Rabbat M., et al. Self-supervised learning from images with a joint-embedding predictive architecture. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; 15619–15629. https://doi.org/10.1109/CVPR52729.2023.01499
16. Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations. In: Daumé H. III, Singh A., editors. Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18. Proceedings of Machine Learning Research. Vol. 119. PMLR; 2020; 1597–1607.
17. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training [Internet]. OpenAI; 2018 [cited 2026 Feb 14]. Available from: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
18. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J., Doran C., Solorio T., editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis (MN): Association for Computational Linguistics; 2019; 4171–4186. https://doi.org/10.18653/v1/N19-1423
19. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv [Preprint]. 2013 Sep 7. https://doi.org/10.48550/arXiv.1301.3781
20. He K., Fan H., Wu Y., Xie S., Girshick R. Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975
21. Grill J.B., Strub F., Altché F., Tallec C., Richemond P., Buchatskaya E., et al. Bootstrap your own latent: a new approach to self-supervised learning. In: Larochelle H., Ranzato M., Hadsell R., Balcan M.F., Lin H., editors. Advances in Neural Information Processing Systems. Vol. 33. Red Hook (NY): Curran Associates, Inc.; 2020; 21271–21284.
22. Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., et al. Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951
23. Balestriero R., LeCun Y. LeJEPA: provable and scalable self-supervised learning without the heuristics. arXiv [Preprint]. 2025 Nov 14. https://doi.org/10.48550/arXiv.2511.08544
24. Picaud G., Chaumont M., Subsol G., Téot L. SSL based encoder pretraining for segmenting a heterogeneous chronic wound image database with few annotations. In: Yap M.H., Kendrick C., Brüngel R., editors. Diabetic Foot Ulcers Grand Challenge. Cham: Springer; 2025. p. 71–80. Lecture Notes in Computer Science; vol. 15335. https://doi.org/10.1007/978-3-031-80871-5_6
25. Oquab M., Darcet T., Moutakanni T., Vo H., Szafraniec M., Khalidov V., et al. DINOv2: learning robust visual features without supervision. arXiv [Preprint]. 2024 Feb 2. https://doi.org/10.48550/arXiv.2304.07193
26. Dhar M.K., Wang C., Patel Y., Zhang T., Niezgoda J., Gopalakrishnan S., et al. Wound tissue segmentation in diabetic foot ulcer images using deep learning: a pilot study. arXiv [Preprint]. 2024 Jun 23. https://doi.org/10.48550/arXiv.2406.16012
27. Lin T.Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 2020; 42(2): 318–327. https://doi.org/10.1109/TPAMI.2018.2858826
28. Kiprono M. WoundNet-Ensemble: a novel IoMT system integrating self-supervised deep learning and multi-model fusion for automated, high-accuracy wound classification and healing progression monitoring. arXiv [Preprint]. 2025 Dec 20. https://doi.org/10.48550/arXiv.2512.18528
29. Filiot A., Ghermi R., Olivier A., Jacob P., Fidon L., Camara A., et al. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv [Preprint]. 2024 Dec 18. https://doi.org/10.1101/2023.07.21.23292757
30. Vu Y.N.T., Wang R., Balachandar N., Liu C., Ng A.Y., Rajpurkar P. MedAug: contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. In: Jung K., Yeung S., Sendak M., Sjoding M., Ranganath R., editors. Proceedings of the 6th Machine Learning for Healthcare Conference; 2021 Aug 6–7. Proceedings of Machine Learning Research. Vol. 149. PMLR; 2021; 755–769.
31. Chen X., Fan H., Girshick R., He K. Improved baselines with momentum contrastive learning. arXiv [Preprint]. 2020 Mar 9. https://doi.org/10.48550/arXiv.2003.04297
32. Sowrirajan H., Yang J., Ng A.Y., Rajpurkar P. MoCo pretraining improves representation and

transferability of chest X-ray models. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning; 2021 Jul 7–9. Proceedings of Machine Learning Research. Vol. 143. PMLR; 2021; 728–744.

33. Zhang Y., Jiang H., Miura Y., Manning C.D., Langlotz C.P. Contrastive learning of medical visual representations from paired images and text. In: Lipton Z., Ranganath R., Sendak M., Sjoding M., Yeung S., (Eds). Proceedings of the 7th Machine Learning for Healthcare Conference; 2022 Aug 5–6. Proceedings of Machine Learning Research. Vol. 182. PMLR; 2022; 2–25.
34. Oota S.R., Rowtula V., Mohammed S., Liu M., Gupta M. WSNet: towards an effective method for wound image segmentation. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023. p. 3233–3242. https://doi.org/10.1109/WACV56688.2023.00325
35. Thomas S. Medetec Medical Images [Internet]. Medetec; [cited 2026 Feb 14]. Available from: https://www.medetec.co.uk/files/medetec-images.html
36. Blanco G., Traina A.J.M., Traina C. Jr., Azevedo-Marques P.M., Jorge A.E.S., de Oliveira D., Bedo M.V.N. A superpixel-driven deep learning approach for the analysis of dermatological wounds. Comput. Methods Programs Biomed. 2020; 183: 105079. https://doi.org/10.1016/j.cmpb.2019.105079
37. tcollyn. Quality of segmentation masks [Internet]. Hugging Face; 2025 Dec 23 [cited 2026 Feb 14]. Available from: https://huggingface.co/datasets/subbareddyoota/wseg_dataset/discussions/2
38. Oota S.R. WSNET: wound image segmentation [software on the Internet]. GitHub; [cited 2026 Feb 14]. Available from: https://github.com/subbareddy248/WSNET
39. Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N., Hornegger J., Wells W.M., Frangi A.F., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
40. Leo'sCode. Wound images segmentation [2760 samples] [dataset on the Internet]. Kaggle; [cited 2026 Feb 14]. Available from: https://www.kaggle.com/datasets/leoscode/wound-segmentation-images