A Systematic Comparison of Pretrained Visual Representations for Dermatologic Concept Encoding

Linda Wermelinger[1,2], Simone Lionetti[2], Nipun Ranasekara[1,2],
Marc Pouly[2], Alexander A. Navarini[1,3]

[1] University of Basel, [2] Lucerne University of Applied Sciences and Arts,
[3] University Hospital of Basel

✉ linda.wermelinger@unibas.ch

Abstract

The Challenge

AI decision support in dermatology is not clinically useful if clinicians cannot understand the model's reasoning. While deep visual representations can encode clinically meaningful concepts, the consistency across architectures and their entanglement with dataset-specific artifacts remain underexplored.

Our Approach: Systematic Comparison

We conducted a systematic comparison of 11 pretrained models (general, SSL, multimodal, and domain-specific) using the SkinCon dataset.

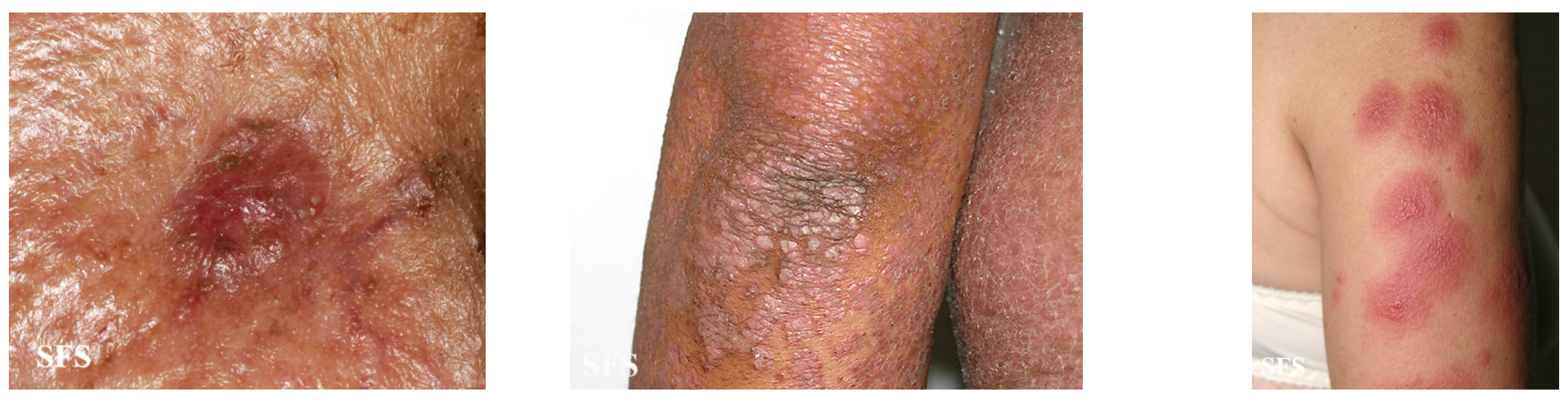
Key Outcomes

- Domain-specific models achieve high concept classification accuracy (e.g., MedSigLIP AUC-ROC: 0.9266).
- Low-dimensional projections show that non-clinical artifacts (e.g., rulers) form stronger clusters than clinical concepts.
- Models are sensitive to dataset shifts (e.g., modality and class imbalance).

Introduction

We utilize SkinCon [1], a multi-label dataset manually annotated by dermatologists. To ensure robust evaluation, we refined the original 48 classes to 31 clinical concepts (e.g., erythema, plaque, scale) by excluding those with fewer than 30 samples. The data was cleaned to remove off-topic images and near-duplicates.

Fig. 1 shows example images from SkinCon with their respective clinical concept annotations.



(a) Papule, Erythema

(b) Papule, Plaque, Xerosis, Scale, Purple

(c) Plaque, Erythema

Figure 1. SkinCon samples [2] annotated with clinical concepts [1], including non-clinical artifacts such as watermarks.

Methods

We primarily evaluate fixed image embeddings from diverse ViT-B/16 architectures to isolate the impact of training objectives from model scale. Performance is assessed through linear probing and two distance-based strategies (Fig. 2), where we compare image features to text descriptions of each concept instead of training a full classifier.

References & Notes

[1] R. Daneshjou, M. Yuksekgonul, Z. R. Cai, R. Novoa, and J. Y. Zou, "SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 18 157–18 167.

[2] M. Groh et al., "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 1820–1828, ISBN: 978-1-6654-4899-4. DOI: 10.1109/CVPRW53098.2021.00201

This research was funded by the Swiss National Science Foundation (SNSF) under grant 20HW-1 228541. Copies of this poster obtained through QR and/or AR codes are for personal use only and may not be reproduced without written permission from the authors. Digital copies of this poster are for personal use only and may not be reproduced without written permission from the authors.

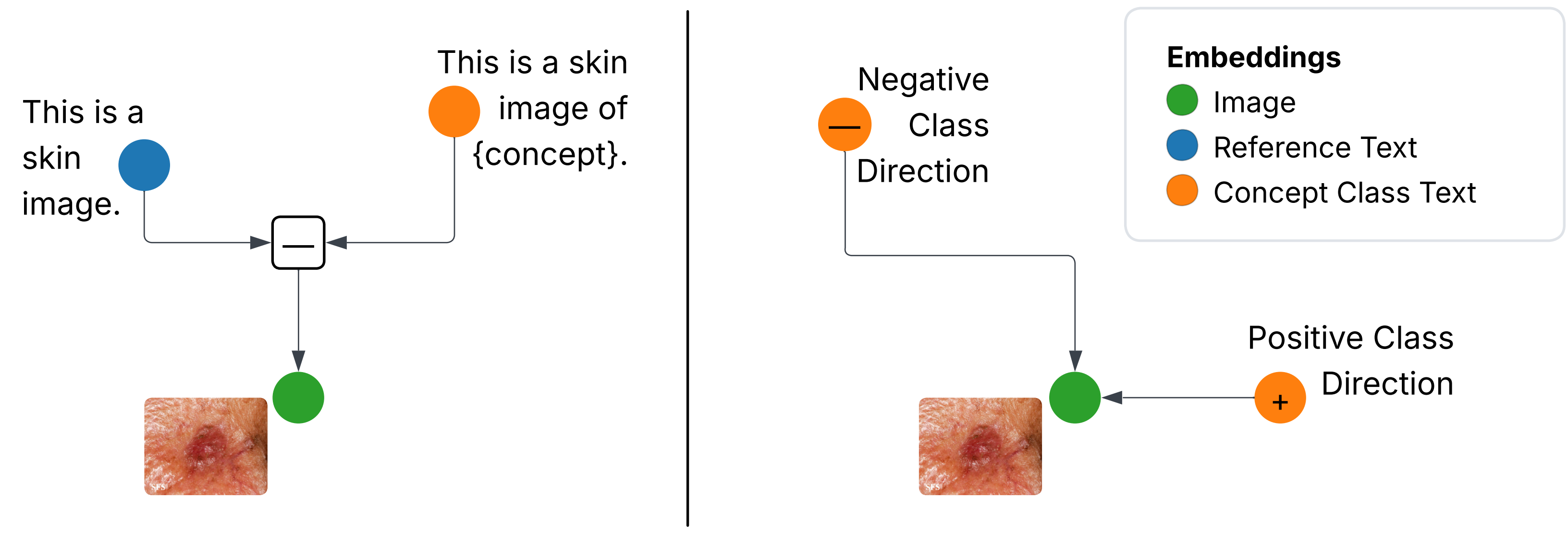


Figure 2. Distance-based Evaluation Strategies. Left (MONET-style): Subtracting reference embeddings from class targets to isolate clinical semantics. Right (DualCoOP-style): Computing similarity between image features and learned positive/negative class prompts. Image from [2].

Experiments

UMAP analysis (Fig. 3) shows that embeddings cluster primarily by dataset source and artifacts (e.g., measurement rulers).
Performance evaluation (Table 1) confirms that domain-specific pretraining improves concept encoding. Among comparable ViT-B/16 backbones, MAKE achieves the highest performance (AUC-ROC: 0.9173). Distance-based evaluation approaches the performance of linear probing. The optimal strategy is model-dependent: DualCoOP yields the best results for MedSigLIP (AUC-ROC: 0.863), while MONET-style subtraction is more effective for MAKE (AUC-ROC: 0.822).

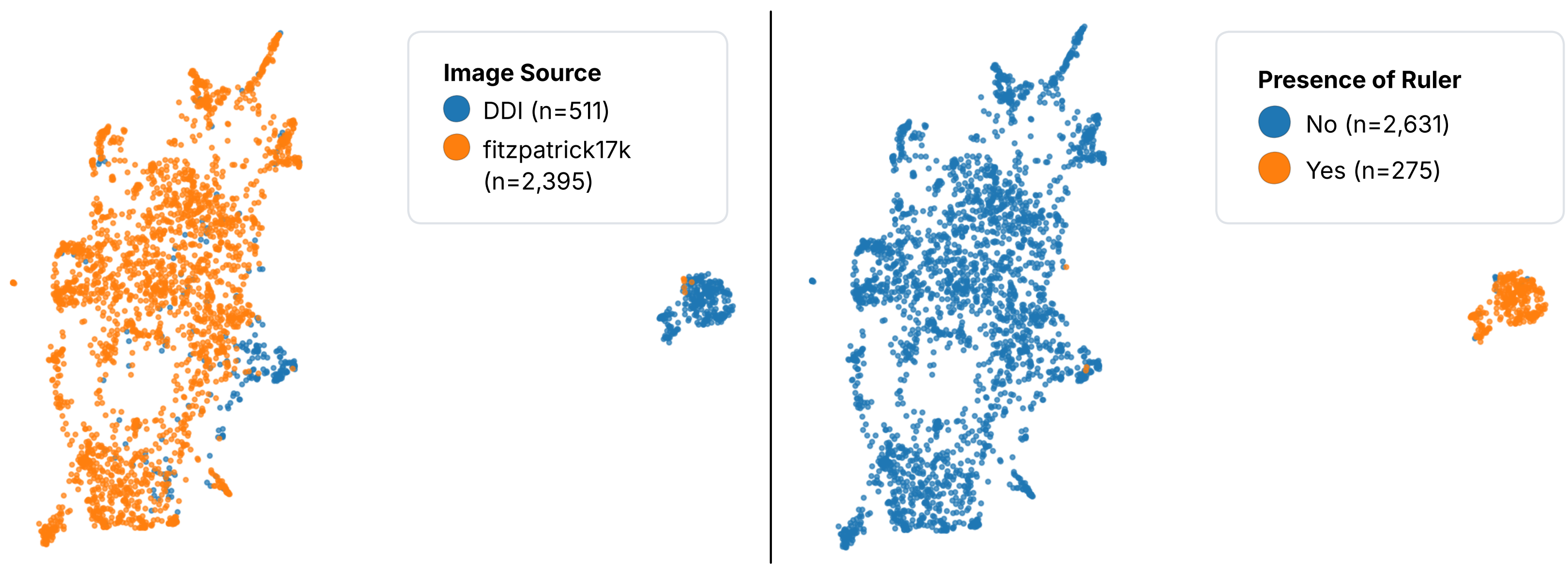


Figure 3. UMAP projections (DINOv3): Embeddings cluster by dataset source (left), with a distinct sub-cluster in DDI (right) aligned specifically with measurement rulers, indicating representation bias toward non-clinical artifacts.

Table 1. Comparison of concept classification with linear probing on SkinCon

	Vision Encoder	Training Paradigm	Val AUC-ROC (Mean ± SD)
ViT-B/16 (baseline)	ViT-B/16	Supervised (ImageNet)	0.7771±0.0013
DINOv3	ViT-B/16, LVD-1689M	Self-Supervised	0.8912±0.0004
DINOv2	ViT-B/14	Self-Supervised	0.8639±0.0014
Derm Foundation	ResNet101x3	Domain-Specific	0.8837±0.0005
PanDerm	ViT-B/16	Domain-Specific	0.8850±0.0001
CLIP	ViT-B/16	Multimodal	0.8769±0.0003
MONET	ViT-L/14	Domain-Specific	0.8822±0.0003
MAKE	ViT-B/16	Domain-Specific	**0.9173±0.0003**
DermLIP	ViT-B/16	Domain-Specific	0.9140±0.0003
SigLIP18	ViT-L/16	Multimodal	0.9147±0.0003
MedSigLIP	So400m	Domain-Specific	**0.9266±0.0001**

Conclusion

While modern vision models successfully encode dermatologic concepts, they remain highly sensitive to domain shifts and non-clinical artifacts, e.g., rulers. Ensuring clinical reliability in explainable AI requires evaluating both quantitative performance and latent representation structure to mitigate dataset-specific biases.

Applied AI Research Lab @ HSLU

Special thanks to Fabian Gröger, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, and Ludovic Amruthalingam for their insightful feedback throughout this project.