Assessing Multimodal Chronic Wound Embeddings with Expert Triplet Agreement

Fabian Kabus[1*], Julia Hindel[2*], Jelena Bratulić[2], Meropi Karakioulaki[3], Ayush Gupta[2], Cristina Has[3], Thomas Brox[2], Abhinav Valada[2], and Harald Binder[1]

[1] Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany
fabian.kabus@uniklinik-freiburg.de (corresponding author)
harald.binder@uniklinik-freiburg.de
[2] Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany
{hindel,bratulic,guptay,brox,valada}@cs.uni-freiburg.de
[3] Department of Dermatology, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany
{meropi.karakioulaki,cristina.has}@uniklinik-freiburg.de

Abstract. Recessive dystrophic epidermolysis bullosa (RDEB) is a rare genetic skin disorder for which clinicians greatly benefit from finding similar cases using images and clinical text. However, off-the-shelf foundation models do not reliably capture clinically meaningful features for this heterogeneous, long-tail disease, and structured measurement of agreement with experts is challenging. To address these gaps, we propose evaluating embedding spaces with expert ordinal comparisons (triplet judgments), which are fast to collect and encode implicit clinical similarity knowledge. We further introduce TriDerm, a multimodal framework that learns interpretable wound representations from small cohorts by integrating wound imagery, boundary masks, and expert reports. On the vision side, TriDerm adapts visual foundation models to RDEB using wound-level attention pooling and non-contrastive representation learning. For text, we prompt large language models with comparison queries and recover medically meaningful representations via soft ordinal embeddings (SOE). We show that visual and textual modalities capture complementary aspects of wound phenotype, and that fusing both modalities yields 73.5% agreement with experts, outperforming the best off-the-shelf single-modality foundation model by over 5.6 percentage points.

1 Introduction

Large-scale dermatology datasets and foundation models have driven substantial progress in skin disease classification and retrieval [14, 15]. However, prior work targets broad taxonomic hierarchies, leaving rare diseases in the long tail of

[*] Fabian Kabus and Julia Hindel contributed equally.
Code: https://github.com/thefabus/triderm

pretraining distributions. Furthermore, current approaches do not capture fine-grained phenotypic structure within specific rare diseases. This is a critical gap, as identifying similar cases and their treatment outcomes is particularly relevant for clinical decision-making in rare diseases.

Recessive dystrophic epidermolysis bullosa (RDEB) exemplifies this challenge. This rare, chronic blistering disease affects approximately 3.5–20.4 per million individuals [2], causing profound skin fragility, chronic non-healing wounds, recurrent infections, and elevated squamous cell carcinoma risk. The typical assessment of patients comprises wound images as well as text descriptions. Yet the substantial heterogeneity in wound appearance and severity remains poorly understood, limiting both clinical decisions and data-driven research.

This challenge motivates embedding-based retrieval, in which wound similarity can be measured directly within a learned representation space. However, validating these similarities requires a structured, expert-anchored assessment to ensure clinical fidelity. Consequently, we introduce triplet judgments of a dermatology expert as a reference for assessing embeddings of image and text data. Given two reference wound images, the medical expert indicates for the third image to which reference image it is most similar in appearance. Unlike categorical labels, triplets encode ordinal similarity structure, representing graded and overlapping phenotypic relationships that resist discrete categorization.

To improve embeddings, we propose TriDerm, which adapts vision foundation models via an attention-pooling module trained with a self-supervised objective on annotated wound regions. For text processing, we propose a synthetic-expert protocol that queries LLMs with wound descriptions and extracts triplet judgments, mirroring the structure of the expert judgment, to adapt LLMs in this small data regime. Since vision and text encoders may capture complementary aspects of wound phenotypes, we also evaluate multimodal fusion strategies that combine their representations.

We make three contributions: an expert-triplet paradigm for evaluating visual, textual, and multimodal representations in a rare-disease setting; wound-level attention pooling with non-contrastive adaptation of visual foundation models; and an SOE approach that recovers clinically meaningful text representations from LLM triplet judgments using only 53 descriptions.

2 Method

2.1 Dataset and Expert Annotations

We collected 53 wound camera images from 21 RDEB patients at a University Medical Center unit specialized on RDEB. A dermatologist manually delineated the individual wounds in each image and provided one free-text description per image characterizing their clinical appearance. This yielded 120 labeled wounds, with images containing between 1–16 wounds across varying sizes and anatomical sites. The images were collected during routine clinical care by treating clinicians using standard camera equipment. As is typical in real-world settings,

no uniform photography protocol (e.g., fixed viewpoint, distance, lighting, or positioning) was mandated, resulting in natural variation in image capture conditions. The text descriptions capture features routinely assessed in wound care, such as granulation tissue level, fibrin coverage, inflammation, depth, erythema, border morphology (hyperkeratotic, fibrotic, scarred), and surface characteristics such as hemorrhagic crusts. In addition, a dermatologist performed triplet judgments: given two reference wound images, the dermatologist indicated which of the reference images a third image is more similar to. This was performed for 513 randomly selected triplets.

2.2 Embedding Methods

Visual Representation: Skin foundation models, DermLIP [14] and DermFM-Zero [15], demonstrate strong capabilities in wound semantics, yet often lack precise differentiation of wound severity and fine-grained degrees of tissue damage. Consequently, we propose a wound-wise representation learning framework as shown in Fig. 1a. Given a camera image of a wound, we first produce two independent augmented views (I, I'). Both views are passed through a frozen DermLIP-ViT-B [14] backbone, yielding feature maps of shape $\mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, C the feature dimension, and H and W the spatial dimensions. Wound-specific feature vectors are then extracted by sampling all spatial locations within the wound region, resulting in a feature tensor of shape $\mathbb{R}^{B\times N\times C}$, where N denotes the number of sampled wound tokens. To aggregate these features, we apply attention pooling per wound, implemented as a two-layer MLP with interleaved tanh activations, producing an attention mask of shape $\mathbb{R}^{B\times N}$. The attended features are subsequently encoded by a lightweight predictor head comprising a single linear layer followed by Layer Normalization to produce the final wound-wise representations F_V and F'_V for I and I', respectively. We train the attention-pooling and predictor heads with the standard VICReg invariance, variance, and covariance terms [1], which enforce augmentation consistency, prevent representational collapse, and decorrelate the embedding dimensions, respectively. We ablate alternative representation-learning objectives in Sec. 3.4. For evaluation, we obtain one feature embedding per image by averaging its wound-wise representations.

Language Representation: Each wound image is accompanied by an expert-written description (T) that characterizes the clinical appearance, including granulation tissue, fibrin coverage, inflammation, depth, and border morphology. These free-text descriptions serve as input for text-based similarity, as shown in Fig. 1b. Rather than directly probing LLM activations—which may not capture clinically meaningful similarity—we treat the LLM as a synthetic expert. Given three wound descriptions (Cases i, j, k), we prompt the model with system prompt P: "Is wound i more similar to wound j or to wound k?" P establishes an RDEB specialist persona to ground the comparison in disease-specific reasoning. We sample triplets uniformly from all possible combinations and collect the

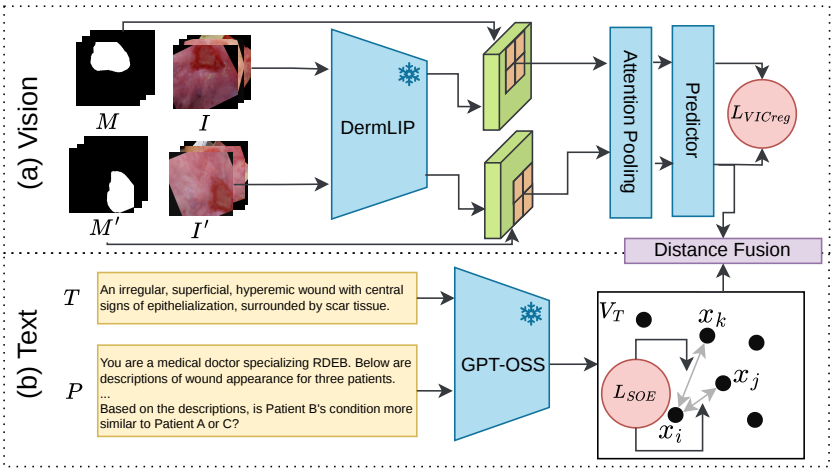

Fig. 1: Overview of TriDerm. (a) In the vision branch, two augmented wound images I, I' and their masks M, M' are encoded by a frozen DermLIP backbone. Attention pooling aggregates the wound-region features, and a predictor is trained with the VICReg objective to produce visual representations F_V. (b) In the text branch, an expert-written wound description T and an RDEB-specific comparison prompt P are passed to a frozen GPT-OSS model. SOE converts the resulting triplet judgments into text representations F_T; V_T denotes the recovered text space, and x_i, x_j, x_k are the embedded cases. TriDerm combines the normalized pairwise distance matrices of F_V and F_T using uncertainty-aware late fusion. Snowflakes denote frozen models.

LLM's binary responses. These synthetic triplet judgments define ordinal constraints on wound similarity. We recover a continuous embedding by fitting Soft Ordinal Embedding (SOE) [12] to the LLM-derived triplets. SOE minimizes a hinge loss that enforces $d(F_T^{(i)}, F_T^{(j)}) < d(F_T^{(i)}, F_T^{(k)})$ when the anchor i is judged closer to j than to k:

$$\mathcal{L}_{\text{SOE}} = \sum_{(i,j,k)} \max\big(0,\ \delta + \|F_T^{(i)} - F_T^{(j)}\| - \|F_T^{(i)} - F_T^{(k)}\|\big), \tag{1}$$

where d denotes Euclidean distance in the SOE constraint and δ is a margin hyperparameter. The resulting embedding F_T captures the LLM's expressed medical reasoning rather than its internal activation geometry.

Distance Fusion: Vision and text embeddings may capture complementary aspects of wounds. We evaluate two fusion strategies that operate on pairwise distance matrices without requiring additional training. Both approaches rely on min-max normalized distance matrices: $\tilde{D}_V$ and $\tilde{D}_T$ denotes the normalized vision and text distance matrices, respectively.

Uncertainty-aware fusion:
Modality reliability varies per sample. For modality $m \in \{V, T\}$, we measure confidence as the variance $\sigma^2_{m,i}$ of distances from sample i to all other samples, where high variance signals clear separation and low variance (equidistance) indicates uncertainty:

$$w_i = \frac{\alpha \cdot \sigma^2_{V,i}}{\alpha \cdot \sigma^2_{V,i} + (1-\alpha) \cdot \sigma^2_{T,i}}, \tag{2}$$

where $\alpha = 0.7$ to account for the vision model's empirically stronger performance. The fused distance estimates per-sample are defined as:

$$D_{\text{fused}}(i,j) = \bar{w}_{ij} \cdot \tilde{D}_V(i,j) + (1 - \bar{w}_{ij}) \cdot \tilde{D}_T(i,j), \quad \bar{w}_{ij} = \tfrac{1}{2}(w_i + w_j). \tag{3}$$

Similarity-based fusion: We convert distances to similarities $S_m = 1 - \tilde{D}_m$ and compute $S_{\text{fused}} = S_V \odot S_T$, where $\odot$ denotes element-wise multiplication, yielding $D_{\text{fused}} = 1 - S_{\text{fused}}$. Since vision and text rely on different features, their errors tend to be weakly correlated. Consequently, agreement filters spurious matches that arise from a single modality.

3 Experiments

3.1 Evaluation Metrics

We evaluate embedding quality via triplet agreements from a dermatologist. Following Sec. 2, we define the triplet margin $\Delta_t = d(i,k) - d(i,j)$, where d is the pairwise distance produced by the evaluated representation or fusion method. A triplet is considered correct when $\Delta_t > 0$ for expert-annotated triplets. To account for unequal triplet counts per anchor due to skipped comparisons, we report anchor-balanced agreement, by weighting the triplets $t \in \mathcal{T}_u$ of each anchor $u \in \mathcal{A}$ equally:

$$\text{Bal. Agr.} = \frac{1}{|\mathcal{A}|} \sum_{u \in \mathcal{A}} \frac{1}{|\mathcal{T}_u|} \sum_{t \in \mathcal{T}_u} \mathbf{1}[\Delta_t > 0]. \tag{4}$$

We also report micro agreement $\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbf{1}[\Delta_t > 0]$, which weights all triplets equally. Treating the prediction as a binary classification task, we compute macro-F1 as the average of the per-class F1 scores, where each F1 is the harmonic mean of precision and recall. Finally, Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$ measures chance-corrected agreement, with p_o the observed micro agreement and p_e the expected agreement under independence of predictions and labels.

3.2 Experimental Settings

For the text modality, we query `gpt-oss-120b` with 68,913 triplets constructed from expert descriptions. We fit SOE embeddings to the resulting constraints using Adam with AMSGrad, learning rate 0.05, batch size 2048, and hinge margin

Table 1: Agreement metrics for text, vision, and fused embeddings. Balanced anchor agreement, micro agreement, and macro-F1 are reported in %; Cohen's κ is unitless.

Modality	Method	Bal. Agr.↑	Micro↑	Macro-F1↑	κ↑
TEXT	MiniLM [13]	56.9	58.0	57.7	0.155
	gpt-oss-120b mean [8]	58.9	58.2	58.0	0.160
	MedGemma-27B mean [10]	61.3	61.8	61.5	0.231
	PubMedBERT [5]	61.7	61.4	61.4	0.230
	RoBERTa [7]	63.7	63.3	63.1	0.262
	TriDerm (Text only)	**67.4**	**66.1**	**66.1**	**0.321**
VISION	MedGemma-27B [10]	50.3	52.4	52.2	0.044
	SigLIP [17]	56.3	58.0	57.9	0.165
	SAM3 [3]	61.9	60.8	60.8	0.218
	PanDerm [16]	58.4	58.0	58.0	0.163
	DINOv3 [11]	66.9	67.2	67.1	0.350
	DermLIP [14]	67.9	63.1	62.9	0.258
	TriDerm (Vision only)	**71.6**	**70.2**	**70.1**	**0.403**
FUSED	TriDerm (similarity)	72.5	72.1	72.1	0.442
	TriDerm (uncertainty)	**73.5**	**72.3**	**72.3**	**0.446**

$\delta = 0$. We train for 50 epochs with anchor-balanced sampling, which resamples triplets each epoch to give equal weight to all anchors regardless of their triplet count. The embedding dimension is set to 4 and we ablate key settings in Sec. 3.4.

The vision model predictor head is trained for 50 epochs with a batch size of 32, a learning rate of 10^{-3}, a weight decay of 10^{-5}, and a cosine learning rate schedule. The VICReg weights for the invariance, variance, and covariance terms are set to their default values of 25, 25, and 1, respectively. The embedding dimension is set to 512. Input images are resized and non-empty randomly cropped to a resolution of 224×224. Data augmentations comprise random horizontal and vertical flips, shifts, scaling, rotations, color jitter, brightness and contrast adjustments, Gaussian noise, and Gaussian blur. At inference time, images are resized to 224×224.

3.3 Results

We compare against relevant baselines for each modality. For text, we consider RoBERTa [7], PubMedBERT [5], MiniLM [13], and MedGemma-27B [10]. For gpt-oss-120b [8] and MedGemma-27B, we create representations using mean pooling over token embeddings. For vision, we benchmark against SigLIP [17], PanDerm [16], SAM3 [3], MedGemma [10], DINOv3 [11], and DermLIP [14]. For all baselines, we use the respective training image size and perform mean pooling of marked wound features.

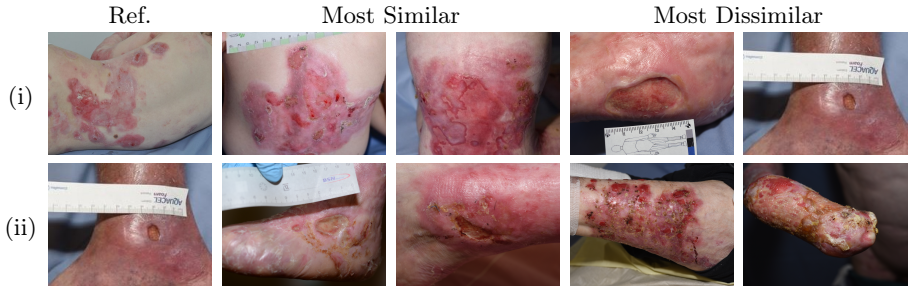

Fig. 2: Nearest neighbor retrieval. In (i), similar wounds show superficial granulating lesions; dissimilar are deep ulcers. In (ii), similar wounds are deep fibrotic ulcers; dissimilar are superficial and inflamed.

Tab. 1 summarizes the performances. For text, SOE yields a gain of 3.7 percentage points (pp) over the strongest baseline (RoBERTa), demonstrating the benefit of encoding medical knowledge into an embedding space via triplet structure. For vision, our proposed wound attention pooling with non-contrastive learning surpasses the best baseline (DermLIP) by 3.7 pp, indicating it captures clinically relevant wound features more effectively. Notably, MedGemma achieves 61.3% on text (above baseline) but only 50.3% on vision (random chance), indicating its visual representations lack discriminative information for this task despite reasonable text performance. Vision and text capture complementary information: uncertainty-aware fusion reaches 73.5%, a gain of 1.9 pp over vision alone (71.6%) and 6.1 pp over text alone (67.4%), consistent with complementary signal from the two modalities. Similarity-based fusion reaches 72.5%, also improving over both unimodal branches. We report balanced anchor agreement, micro agreement, macro-F1, and Cohen's κ, which handle class imbalance and chance agreement differently. All metrics yield comparable rankings, indicating that the comparison is not driven by a single treatment of class imbalance or chance agreement. On the Landis–Koch scale, fusion achieves $\kappa = 0.45$ (moderate agreement), compared to $\kappa \approx 0.32$–0.40 for unimodal methods.

Fig. 2 shows that nearest neighbors share clinically relevant patterns in wound depth, granulation, and inflammation; the selected similar and dissimilar examples have $d < 0.35$ and $d > 0.8$, respectively.

3.4 Ablation

Tab. 2a reports SOE ablations over embedding dimensionality, triplet budget and oracle models. Embedding dimensionality exhibits an optimum at $D = 4$ (67.4%), with performance degrading for both lower and higher dimensions. Very low-dimensional embeddings ($D = 2$) lack the capacity to capture the underlying similarity structure, while higher-dimensional spaces overfit to noise in the LLM's triplet judgments. Regarding the triplet budget, agreement improves

Embedding Dim.				
2	3	4	5	6
62.6	65.0	**67.4**	64.2	64.9
No. Triplets				
0.1%	1%	10%	100%	
55.2	62.6	63.6	**67.4**	
LLM Oracle				
MedG	GPT-20	GPT-120		
61.1	61.7	**67.4**		

(a) Text adaptation

Feature Pooling	
Mean	69.3
Att. (ours)	**71.6**
Loss Function	
Contrastive [6]	68.1
Triplet [9]	70.1
VICReg (ours)	**71.6**

(b) Vision adaptation

Table 2: Ablation studies. (a) TriDerm SOE ablations using GPT-120B as oracle. (b) Vision model pooling and loss variations. Bal. Agr. is reported in %.

monotonically from 55.2% at 0.1% of triplets to 67.4% with the full set, indicating that using all triplet constraints helps uncover the LLM's latent similarity structure. At very small budgets (0.1%), agreement degrades to slightly above chance (50%), confirming the need for sampling a larger subspace of triplet questions. We also compare different LLMs as triplet oracles: gpt-oss-120b (67.4%) outperforms both gpt-oss-20b (61.7%) and the medical-domain MedGemma-27B (61.1%), which also underperforms mean token embeddings (61.3%). Among the three evaluated LLMs, the general-purpose model performs best on pairwise similarity judgments.

Tab. 2b reports the effect of different self-supervised adaptation strategies. Attention pooling over wound regions outperforms mean pooling with the same predictor head by 2.3 pp, because it more effectively weights characteristic wound features. Regarding different self-supervised loss functions, the triplet loss achieves its best performance with a batch size of 8, while contrastive learning benefits from a larger batch size of 128 and a higher learning rate of 0.001, consistent with the known sensitivity of contrastive objectives to the number of in-batch negatives [4]. Finally, the non-contrastive method VICReg [1] outperforms both alternatives under this low-data regime, which we attribute to its independence from negative pairs—a property that confers an advantage when the training set is small.

4 Conclusion

Our findings suggest that visual and text foundation models encode some clinically relevant structure, but miss fine-grained distinctions for RDEB wounds which experts consider important. Targeted adaptation, i.e. VICReg fine-tuning for vision and soft ordinal embedding for LLM triplet judgments, improves this

alignment using only this 53-image cohort. A gap with expert judgment remains, yet multimodal fusion narrows this discrepancy by combining complementary information from vision and text. Our proposed triplet-based evaluation is effective when labeled data is scarce but expert similarity judgments are obtainable.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB 1597.

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

References

1. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (ICLR) (2022)
2. Bardhan, A., Bruckner-Tuderman, L., Chapple, I.L.C., Fine, J.D., Harper, N., Has, C., Magin, T.M., Marinkovich, M.P., Marshall, J.F., McGrath, J.A., Mellerio, J.E., Polson, R., Heagerty, A.H.: Epidermolysis bullosa. Nature Reviews Disease Primers **6**(1), 78 (Sep 2020). https://doi.org/10.1038/s41572-020-0210-0
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 1597–1607. PMLR (2020), https://proceedings.mlr.press/v119/chen20j.html
5. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1) (2021). https://doi.org/10.1145/3458754
6. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692
8. OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., et al.: gpt-oss-120b & gpt-oss-20b model card (Aug 2025). https://doi.org/10.48550/arXiv.2508.10925
9. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)

10. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
11. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
12. Terada, Y., Luxburg, U.V.: Local Ordinal Embedding. In: International Conference on Machine Learning (Jun 2014)
13. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Advances in Neural Information Processing Systems (2020), https://arxiv.org/abs/2002.10957, arXiv:2002.10957
14. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H., Tschandl, P., Kittler, H., Ge, Z.: Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology (Apr 2025). https://doi.org/10.48550/arXiv.2503.14911
15. Yan, S., Li, X., Mo, D., Tschandl, P., Jiang, Y., Wang, Z., Hu, M., Ju, L., Vico-Alonso, C., Zheng, Y., Liu, J., Zhou, J., Chello, C., Cheung, J.G., Anriot, J., Thomas, L., Primiero, C., Tan, G., Ng, A.B., See, S., Tang, X., Ip, A., Liao, X., Bowling, A., Haskett, M., Zhao, S., Janda, M., Soyer, H.P., Mar, V., Kittler, H., Ge, Z.: A Vision-Language Foundation Model for Zero-shot Clinical Collaboration and Automated Concept Discovery in Dermatology (Feb 2026). https://doi.org/10.48550/arXiv.2602.10624
16. Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., et al.: A multimodal vision foundation model for clinical dermatology. Nature Medicine pp. 1–12 (2025)
17. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11941–11952. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01100