Enhancing Skin Disease Diagnosis: Interpretable Visual Concept Discovery with SAM*

Xin Hu[1,†], Janet Wang[1,†], Jihun Hamm[1], Rie R Yotsu[2], Zhengming Ding[1,‡]
[1]Department of Computer Science, Tulane University
[2]Department of Tropical Medicine, Tulane University

Abstract

Current AI-assisted skin image diagnosis has achieved dermatologist-level performance in classifying skin cancer, driven by rapid advancements in deep learning architectures. However, unlike traditional vision tasks, skin images in general present unique challenges due to the limited availability of well-annotated datasets, complex variations in conditions, and the necessity for detailed interpretations to ensure patient safety. Previous segmentation methods have sought to reduce image noise and enhance diagnostic performance, but these techniques require fine-grained, pixel-level ground truth masks for training. In contrast, with the rise of foundation models, the Segment Anything Model (SAM) has been introduced to facilitate promptable segmentation, enabling the automation of the segmentation process with simple yet effective prompts. Efforts applying SAM predominantly focus on dermatoscopy images, which present more easily identifiable lesion boundaries than clinical photos taken with smartphones. This limitation constrains the practicality of these approaches to real-world applications. To overcome the challenges posed by noisy clinical photos acquired via non-standardized protocols and to improve diagnostic accessibility, we propose a novel Cross-Attentive Fusion framework for interpretable skin lesion diagnosis. Our method leverages SAM to generate visual concepts for skin diseases using prompts, integrating local visual concepts with global image features to enhance model performance. Extensive evaluation on two skin disease datasets demonstrates our proposed method's effectiveness on lesion diagnosis and interpretability.

1. Introduction

Current AI-assisted diagnostic systems demonstrate expert-level capability in classifying skin cancers, which are
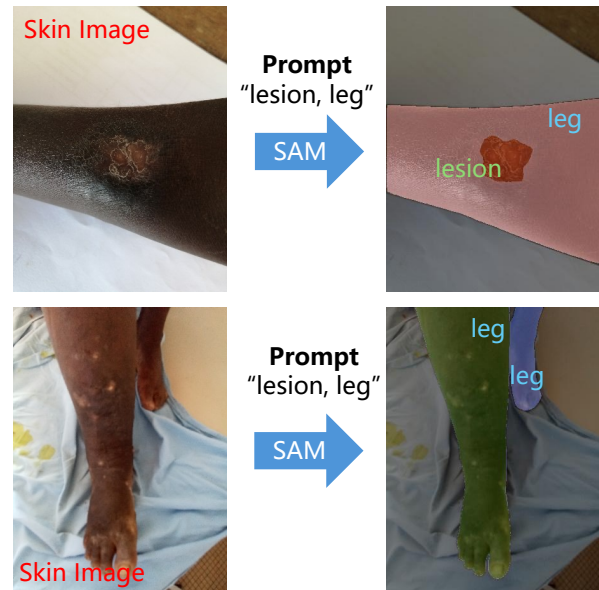
Figure 1. Generated masks on two skin samples with prompt "lesion" and "leg". The first row shows a "Buruli ulcer" image, in which the lesion part is clear to be detected as the visual concept, while in the second row - "Mycetoma", the lesion boundary is ambiguous to be recognized.

often identified visually [3, 4, 9, 26, 30, 33, 34]. Given that early and accurate diagnosis is important in improving treatment outcomes, these systems can significantly contribute to teledermatology as diagnostic and decision-support tools [10]. By utilizing photos captured from portable devices like smartphones, these methods promote diagnostic accessibility in rural areas [6]. However, such systems are susceptible to under-diagnosis due to non-standardized acquisition environments and protocols. Therefore, the model's ability to accurately recognize the region of interest (ROI) within noisy backgrounds is crucial for precise diagnosis.

While image segmentation techniques such as MaskR-CNN [16], DeepLab [5], and Panoptic Segmentation [22] can be employed to localize ROIs and enhance diagnostic accuracy, they rely on fine-grained bounding box or pixel-

wise semantic annotations. To acquire such annotations from manual labeling is labor-intensive and impractical, especially when dealing with noisy images. The Segment Anything Model (SAM) [23] can serve as a potential solution, capable of automating the generation of two-level masks: one for the region of the body and another for the lesion. These masks also enable the explainable investigation of different features in the diagnostic results.

Two works have explored SAM to locate skin lesions and automate the segmentation [18, 19]. Both studies primarily focus on the well-curated HAM10000 Dataset, which includes 10,015 verified dermoscopy images and groundtruth segmentation masks [32, 33]. In this dataset, examples are abundant, lesion boundaries are clear, and background noise is minimized. In contrast, clinical photos collected in real-world applications present greater difficulty. The contrast between lesions and backgrounds in these photos is often less distinct due to factors such as (1) a wide variance in lesion presentation, including differences in shape, texture, color, and location, and (2) inconsistent image quality caused by varying lighting, distance, angles, environments, or image resolutions. These variations make it difficult for the model to accurately identify the ROI from the background, limiting the practicality of existing methods. Moreover, clinical photos of skin conditions are often limited in size. Therefore, the potential application of SAM in small datasets of clinical photos remains under-explored. Figure 1 shows results on example images using SAM.

In this work, we propose an innovative framework to locate the region of interest and diagnose skin diseases in the clinical photos. Empowered by SAM, our framework generates two-level masks to enhance the accuracy of AI diagnoses and provide robust support for remote medical personnel, enhancing diagnostic reliability and transparency. To sum up, we highlight our contributions as follows:

- We leverage the capabilities of a foundational segmentation AI model to generate two-level masks over the body and lesions with prompts, which delineates the regions of interest and reveals potential visual concepts essential for skin diagnosis.

- We introduce a novel Cross-Attentive Fusion framework that integrates knowledge from global image features and local visual concepts to enhance the model's diagnostic accuracy. This framework also facilitates interpretable analysis of visual concepts crucial for accurate diagnosis.

- We comprehensively evaluate our method on two distinct skin image datasets of challenging clinical photos. Through detailed interpretation and analysis, our method demonstrates robustness and effectiveness in real-world scenarios, elevating confidence in its diagnostic capabilities.

2. Related Works

2.1. Visual Segment Generation

Image segmentation techniques such as MaskRCNN [16], DeepLab [5], and Panoptic Segmentation [22] are effective for localizing regions of interest and enhancing model accuracy, but they typically require fine-grained bounding box or pixel-wise semantic annotations. With a growing focus on scalability, pre-trained foundational models have gained prominence in machine learning, serving as robust starting points for various downstream tasks [31]. Responding to this trend, the Segment Anything Model (SAM) has emerged, pushing image segmentation into the realm of foundational models.

SAM is a promptable segmentation model pre-trained on SA-1B, a vast dataset containing over 1 billion masks derived from 11 million licensed and privacy-preserving images. This extensive training ensures strong generalization across diverse data distributions. SAM supports flexible prompts such as single points, sets of points, bounding boxes, or text. Its architecture is elegantly simple yet effective: a powerful image encoder computes an image embedding, a prompt encoder embeds prompts, and these inputs are fused in a lightweight mask decoder to predict segmentation masks. Specifically, SAM's image encoder utilizes a Masked Autoencoder (MAE) [15] pre-trained on a Vision Transformer (ViT) [8]. This adaptation allows it to handle high-resolution inputs while capturing fine-grained details and long-range dependencies. The prompt encoder converts prompts into fixed-length embeddings that capture semantic meanings. These embeddings are combined with image encoder outputs to generate a set of feature maps used by the mask decoder to produce segmentation masks.

2.2. Skin Lesion Analysis

Skin lesion images are typically captured in two forms: dermatoscopy images and clinical photos. Dermoscopy images are close-up views of pigmented skin lesions captured by using professional microscopy, a dermatoscope. These images focus on the lesion, typically excluding any background, resulting in a uniform and consistent visual presentation. Clinical photos, on the other hand, are often taken with portable devices like smartphones and contain more noisy backgrounds. Extensive research has explored the use of deep Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [12, 27] for skin lesion segmentation to standardize lesion variations and enhance diagnostic accuracy [14, 21, 24, 28, 36, 39]. CNN-based methods for skin lesion segmentation typically employ supervised learning with large labeled datasets to extract spatial features and semantic maps from images. In contrast, GAN-based approaches address data scarcity through unsupervised learning but still require fine-grained groundtruth
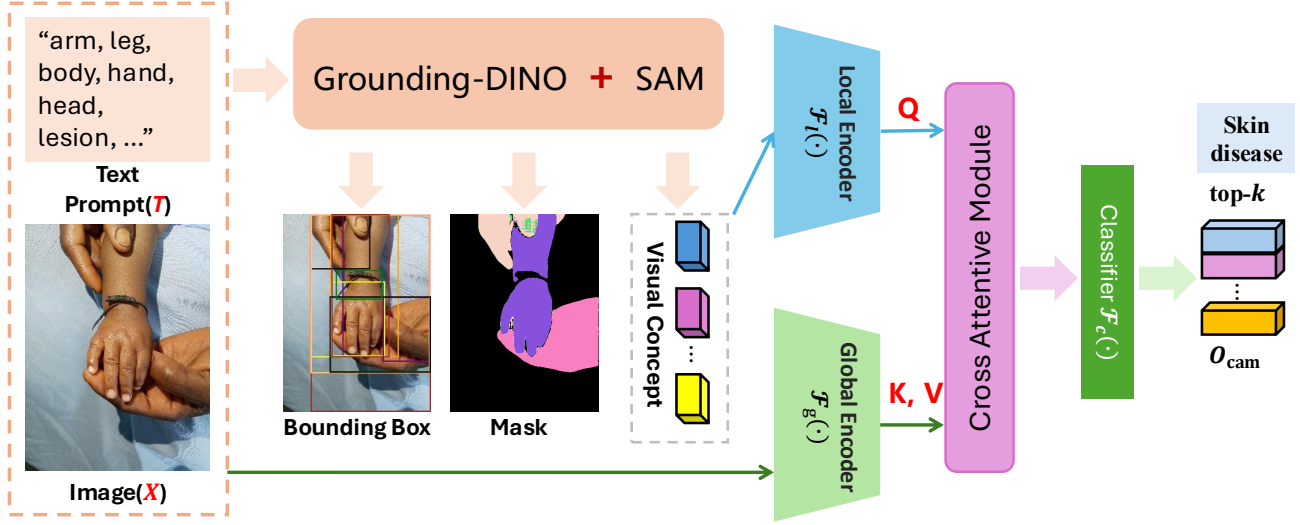
Figure 2. Overall framework of our proposed model, where we use Grounding-DINO and SAM to extract visual concepts, bounding boxes, and segmentation masks. The local encoder $\mathcal{F}_l(\cdot)$ converts visual concepts to local tokens and sets them as "query" prompts to trigger the salient area of the encoded global image with cross attentive module. The classifier $\mathcal{F}_c(\cdot)$ transfers the latent features to CAM for classification and interpretation for the decision-making process.

masks, often unavailable in real-world applications.

With the emergence of large foundation models, there has been an effort to utilize the Segment Anything Model (SAM) for dermatoscopy images, using simple prompts [18, 19]. Specifically, [19] fine-tuned SAM on the HAM10000 Dataset with fine-grained groundtruth masks, achieving state-of-the-art segmentation performance with the guidance of prompts. [18] utilized SAM to segment cancerous areas in HAM10000 images before feeding the segmented images to a pre-trained Vision Transformer (ViT) model for classification. Their results demonstrated ViT's superiority over traditional architectures. Despite these advancements, adapting SAM to clinical data captured by portable devices poses challenges due to limited dataset sizes for fine-tuning and the complexity of lesion manifestations for off-the-shelf SAM applications [40]. Therefore, the potential of SAM in low-data scenarios and with clinical photos remains to be fully explored.

3. Methods

3.1. Preliminary and Motivation

Our objective is to develop an accurate and interpretable diagnostic model for complex image data. Given a batch of skin disease images, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_i, \ldots, x_N\}$, where N is the total number of images, the model aims to predict the most likely skin disease for each image, represented as $\{\mathbf{y}_i\}_{i=1}^N$. Here, $\mathbf{y}_i \in \mathbb{R}^C$, with C representing the total number of condition categories, and $\mathbf{y}_{i,c}$ the presence of skin disease category c in $\mathbf{x}_i$.

Considering the skin images are usually captured from smartphones and tablets in the field, they lack standardiza-

tion and often contain confounding background noise. The ability to easily discern the region of interest (ROI) or lesion in a poor-quality or noisy image is crucial to accurate diagnosis. In this study, we will explore the Segment Anything Model (SAM) [23] to detect visual concepts with specific prompts, then develop a cross-attentive fusion model by leveraging the global image information and local visual cues to improve the skin disease diagnosis performance.

3.2. Visual Concepts Identification

Due to the often complicated background noise, end-to-end deep learning is not sufficient to fully extract informative features from raw images with disease labels. Though SAM can capture local visual concepts of lesions, the original SAM needs bounding boxes or coarse masks as input and might produce features containing background noise. Thus, we modify variants of SAM in public repository [2] which utilizes Grounding DINO [25], to obtain visual concepts and provide related bounding boxes.

As shown in Figure 2, we use keywords describing human body parts and "lesion" as text prompts and input them with the original image to the SAM variant to generate bounding boxes, visual concepts, and segmentation masks. Figure 1 lists some detected visual concepts. It is noted that the text prompt "lesion" might correctly localize the pathological area for certain skin diseases whose lesion boundaries are clear, but it is challenging for other complicated conditions, such as "Mycetoma". Still, these captured visual concepts can offer extra information and help us identify the relevant segments of the body or the lesion.

We define local visual concepts as $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\} \in \mathbf{V}$. By focusing on these localized visual elements, our model

can diagnose skin diseases more accurately and effectively. Additionally, given the visual concepts identified by SAM, we can enhance the interpretability of our model by pinpointing the specific local patterns that lead to a prediction. The framework provides valuable insights into the decision-making process, thereby increasing its transparency and trustworthiness as a diagnostic system. This dual focus on accuracy and interpretability is crucial for developing a reliable diagnostic tool that can assist healthcare professionals in making informed decisions.

3.3. Cross-Attentive Fusion Model

Due to the unreliability of certain visual concepts from SAM (Figure 1), we concurrently leverage both global image features and local visual concepts. As illustrated in Figure 2, our proposed dual-branch framework is designed to harness the complementary strengths of these two sources of information. This approach allows for a more robust and comprehensive understanding of the diseases by integrating detailed local visual concepts with overarching global image features, thereby improving the overall accuracy and reliability of the analysis.

Firstly, we introduce a global feature encoder, denoted as $\mathcal{F}_g(\cdot)$, to extract the global features $\mathbf{Z}_g \in \mathbb{R}^{N \times 1 \times D}$ for the entire image. This is based on the premise that the overall image encapsulates most of the essential information. However, it is crucial to pinpoint the specific location of the skin disease's origin. Moreover, global features often encompass background noise, which can degrade the prediction performance. Thus, while global features provide a broad context, their inherent noise necessitates complementary local analysis for precise disease localization and classification.

To address this issue, we propose exploring local visual concepts that potentially preserve more focused information about skin disease, thereby enhancing model decision-making. As illustrated in Figure 2, we input the local visual concepts $\mathbf{V}$ into the local encoder $\mathcal{F}_l(\cdot)$ to extract local features $\mathbf{Z}_l \in \mathbb{R}^{N \times n \times D}$, where n represents the total number of visual concepts. This method effectively filters out background noise and irrelevant objects, ensuring the model focuses on the most severely affected and informative pixels.

To efficiently leverage the complementary information between local concepts and global images, we propose a cross-attentive module where local concepts serve as "query" prompts to highlight the most salient areas within the global features. Specifically, the attention map $\mathbf{M} \in \mathbb{R}^{N \times n \times 1}$ acts as a soft mask between the local concept features $\mathbf{Z}_l$ and the global features $\mathbf{Z}_g$. This attention map is normalized using a global softmax function to capture the most relevant segments between the local and global features. This component further strengthens the model's ability to focus on the critical areas, thereby enhancing the overall precision and effectiveness of the diagnosis system.

Following the Transformer design, we define three variables—query, key, and value—as $\mathbf{Q} = \mathbf{Z}_l \mathbf{W}_q$, $\mathbf{K} = \mathbf{Z}_g \mathbf{W}_k$, and $\mathbf{V} = \mathbf{Z}_g \mathbf{W}_v$, where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the linear projection matrices in $\mathbb{R}^{D \times D}$ used to generate the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$, respectively. These maps are then processed by each Transformer module to generate refined latent features, as described as follows:

$$\begin{cases} \mathbf{I} = \mathbf{Z}_g + \mathcal{F}_{\text{Drop}}(\mathbf{MVW}_{proj}), \\ \mathbf{O} = \mathbf{I} + \mathcal{F}_{\text{MLP}}(\mathcal{F}_{\text{LN}}(\mathbf{I})), \end{cases} \quad (1)$$

where $\mathbf{W}_{proj} \in \mathbb{R}^{D \times D}$ are learnable projection matrices, $\mathcal{F}_{\text{LN}}(\cdot)$ is the layer normalization function, $\mathcal{F}_{\text{Drop}}(\cdot)$ is the dropout module, and $\mathcal{F}_{\text{MLP}}(\cdot)$ is the multi-layer perceptron (MLP) module. The intermediate latent feature is denoted as $\mathbf{I}$. The output $\mathbf{O} \in \mathbb{R}^{N \times n \times D}$, which merges information from both local and global features, is more representative of the input image. To build the cross-attentive modules, we define $\mathbf{M} = \frac{\mathbf{QK}^\top}{\sqrt{n}}$. Additionally, multi-head attention is introduced to expand the capacity of the attention modules, allowing for a richer and more nuanced representation of the relevant features. This enhanced structure ensures a more effective integration and utilization of both local and global information.

3.4. Interpretable Skin Diagnosis

Having obtained the refined latent features $\mathbf{O}$, it is essential to identify which parts contribute the most to future observations in the medical field. Typically, most datasets lack such detailed labels because annotating them is labor-intensive and usually requires professional expertise to ensure label quality in medical research. Inspired by the weakly supervised setting described by [20], we introduce the Class Activation Map (CAM) and develop multi-instance learning (MIL) loss for the final prediction, instead of relying on traditional classification loss.

In our framework, we utilize a classifier $\mathcal{F}_c(\cdot)$, as shown in Figure 2, to transform the latent features $\mathbf{O}$ into $\mathbf{O}_{cam} \in \mathbb{R}^{N \times n \times C}$. Subsequently, a top-$k$ strategy is employed to identify the concepts that contribute the most, denoted $\hat{\mathbf{O}}_{cam} \in \mathbb{R}^{N \times k \times C}$. The final prediction is based on the average value of these top-k concepts, as given by:

$$\mathbf{O}_{\text{pred}} = \mathcal{F}_{\text{AVG}}(\hat{\mathbf{O}}_{\text{cam}}), \quad (2)$$

where $\mathcal{F}_{\text{AVG}}(\cdot)$ represents the average pooling function along the dimension k. The final loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{MIL}}(\mathbf{O}_{\text{pred}}, \mathbf{y}_i), \quad (3)$$

where $\mathcal{L}_{\text{MIL}}(\cdot)$ denotes the multi-instance learning loss function. This approach allows us to effectively utilize the weakly supervised signals to identify and focus on the most

Table 1. Class distribution for the experimental dataset. "AC Dermatitis" is short for Allergic Contact Dermatitis.

NTD	Buruli ulcer	Yaws	Scabies	Leprosy	Mycetoma	Total
Patient #	300	223	160	57	18	758
Image #	787	375	391	135	43	1,731
SCIN	Eczema	AC Dermatitis	Urticaria	Insect Bite	Psoriasis	Total
Patient #	355	214	163	128	61	921
Image #	801	467	338	290	126	2,022

critical parts of the input image, thereby enhancing the accuracy and reliability of the final medical diagnosis.

Equation (2) illustrates that the final decision is based on the average value of the top-k visual concepts, indicating that these k concepts provide the most significant contributions. However, in the context of a skin disease diagnosis system, it is common to search for the final interpretation with one most important local visual concept. As depicted in Figure 6, our framework first determines the most probable skin disease from top-k concepts. Subsequently, the single concept that contributes the most is used as the basis for interpreting our model's decision. This ensures that the diagnosis is not only accurate but also interpretable, allowing medical professionals to understand which specific feature the model deems most indicative of the disease.

4. Experiments

4.1. Datasets

We evaluated our methods using two datasets: MIND-the-SKIN [1] and SCIN [35]. The MIND-the-SKIN project aims to address challenges in Neglected Tropical Diseases (NTDs), a diverse group of skin conditions prevalent in impoverished tropical communities, affecting over 1 billion people. A crucial aspect of the project involves data collection in rural West Africa using portable devices and the development of AI-based diagnostic tools. For our evaluation, we utilized a subset of this dataset comprising 1,731 clinical photos representing five common NTD conditions: leprosy, Buruli ulcers, yaws, scabies, and mycetoma.

Existing skin disease datasets featuring clinical photos [7, 11, 13, 29, 38] are pre-processed to center at lesions and have reduced background noise. In contrast, the SCIN dataset retains its original noise and more closely mirrors real-world inputs, making it a valuable resource for our investigation. The SCIN dataset was collected through a voluntary image donation platform from Google Search users in the United States, with each case including up to three images, all diagnosed by up to three dermatologists. This process results in a weighted skin condition label for each case. To ensure label accuracy, we selected the condition with the highest weight as the final label, excluding ambiguous cases where multiple conditions had equal probabilities. Additionally, to facilitate reliable evaluation and maintain consistency with the NTD dataset in terms of the number of

conditions, we focused on the five largest classes. In our experiments, we used a random split by cases for both datasets to prevent data leakage. Dataset details and class distributions are provided in Table 1.

4.2. Implementation Details

We use the pretrained ViT [8] as the backbone to extract global features from each input image. The global feature extracted by ViT, combined with our top-k mechanism, serves as the baseline for comparison. ViT produces a 768-dimensional feature map with a resolution of 14×14, which we pass through a linear layer to reduce the dimensionality to 256 for further refinement. For the local encoder, we extract visual concepts using our SAM variant, generating 256-dimensional vectors. These local features are refined with two convolutional layers, also outputting 256-dimensional vectors. To maintain consistent input, we fix the number of visual concepts at 30 for the NTD dataset and 25 for the SCIN dataset, based on the maximum number of captured concepts. For samples with fewer visual concepts, we use a random perturbation padding method to ensure consistency in input size. In the cross-attentive module, we implement a multi-head transformer with 8 heads, where each transformer module consists of 1 attention layer with a dimensionality of 256. The classification module includes 3 convolutional layers, with two dropout layers (dropout rate of 0.7) between them to regularize intermediate features.

For hyperparameter choice, we train the whole model with a learning rate of 1e-4 using the Adam optimizer, and the batch size is set as 128 for training and 1 for testing. Considering the overfitting problem, we train all model parts simultaneously with 30 epochs. We choose $k = 5$ for the NTD dataset and 2 for the SCIN dataset. All experiments are conducted with one RTX 4090 GPU.

4.3. Comparison Results

We first tested our proposed method on the NTD dataset from the MIND-the-SKIN Project. As shown in Table 2, we conducted five random splits of cases into training and validation sets and compared our method with three other competitive methods. An immediate observation is that our method consistently outperforms the others in terms of classification accuracy for all splits, validating the effectiveness of leveraging both global and local features for skin lesion

Table 2. Comparison of different methods on the MIND-the-SKIN dataset.

Split Ratio	Method	Precision	Recall	F1	Accuracy
train/total = 0.1	ResNet [17]	0.381	0.324	0.350	0.535
	ViT [8]	**0.393**	0.364	0.378	0.578
	Baseline(global-only)	0.313	0.330	0.321	0.539
	Ours	0.375	**0.381**	**0.378**	**0.579**
train/total = 0.3	ResNet [17]	0.575	**0.538**	0.556	0.721
	ViT [8]	0.603	0.536	0.567	0.746
	Baseline (global-only)	0.606	0.495	0.545	0.743
	Ours	**0.643**	0.509	**0.568**	**0.750**
train/total = 0.5	ResNet [17]	0.623	0.547	0.582	0.777
	ViT [8]	0.554	0.567	0.560	0.760
	Baseline (global-only)	0.550	0.565	0.557	0.764
	Ours	**0.651**	**0.631**	**0.641**	**0.809**
train/total = 0.7	ResNet [17]	0.740	0.588	0.655	0.798
	ViT [8]	0.756	0.663	0.706	0.789
	Baseline (global-only)	0.738	0.671	0.702	0.816
	Ours	**0.805**	**0.768**	**0.786**	**0.847**
train/total = 0.9	ResNet [17]	0.739	0.732	0.735	0.829
	ViT [8]	0.824	0.775	0.799	0.853
	Baseline (global-only)	0.766	0.682	0.722	0.842
	Ours	**0.841**	**0.798**	**0.819**	**0.867**
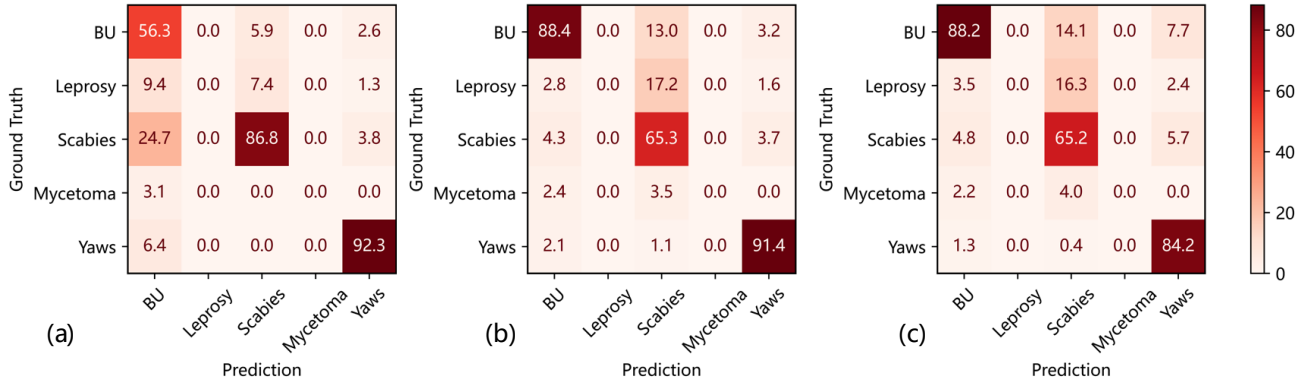
Figure 3. Confusion matrix with different top-k strategies on the MIND-the-SKIN dataset. **(a)** is the confusion matrix of top-1; **(b)** shows the confusion matrix of top-5; **(c)** represents the confusion matrix of top-15.

images. Notably, as the size of the training set increases, the margin of improvement gained by our method over the others also enlarges across the four metrics. This suggests that our method is more sensitive to new information than its competitors and can make full use of additional features to achieve superior performance. Figure 3 reports the confusion matrix when we select different k and train/total = 0.5, which aims to show the category-specific accuracy. Interestingly, we notice our model fails to recognize "Leprosy" and "Mycetoma". The reason is that the lesion is not visually clear for those two diseases and they have the smallest sample sizes, thus easily misclassified. Figure 4 indicates the comparison of each category in the MIND-the-SKIN

dataset. It is observed that though our fusion model shows a little performance degradation in the "Leprosy" disease, it shows an obvious boost in the "Scabies" disease. On average, our proposed fusion method achieves better performance than the baseline method.

Similarly, we performed five random splits for the SCIN dataset in Table 3, as no official splits are available. We observed a similar trend in the SCIN dataset, where our proposed method generally outperforms the baseline. However, we also observed that the overall performance on the SCIN dataset was suboptimal. To the best of our knowledge, no prior studies have experimented with the SCIN dataset, leaving us no references for comparison. One pos-

Table 3. Comparison of different methods on the SCIN dataset.

| Split Ratio	Method	Precision	Recall	F1	Accuracy
train/total = 0.1	Baseline (global-only)	0.196	0.204	0.200	0.385
	Ours	**0.233**	**0.209**	**0.220**	**0.393**
train/total = 0.3	Baseline (global-only)	0.070	0.199	0.104	0.394
	Ours	**0.190**	**0.221**	**0.204**	**0.395**
train/total = 0.5	Baseline (global-only)	0.145	**0.209**	0.171	0.395
	Ours	**0.155**	0.208	**0.177**	**0.397**
train/total = 0.7	Baseline (global-only)	0.207	0.249	0.226	**0.403**
	Ours	**0.275**	**0.255**	**0.265**	0.394
train/total = 0.9	Baseline (global-only)	0.325	0.307	0.316	0.436
	Ours	**0.347**	**0.319**	**0.332**	**0.439**
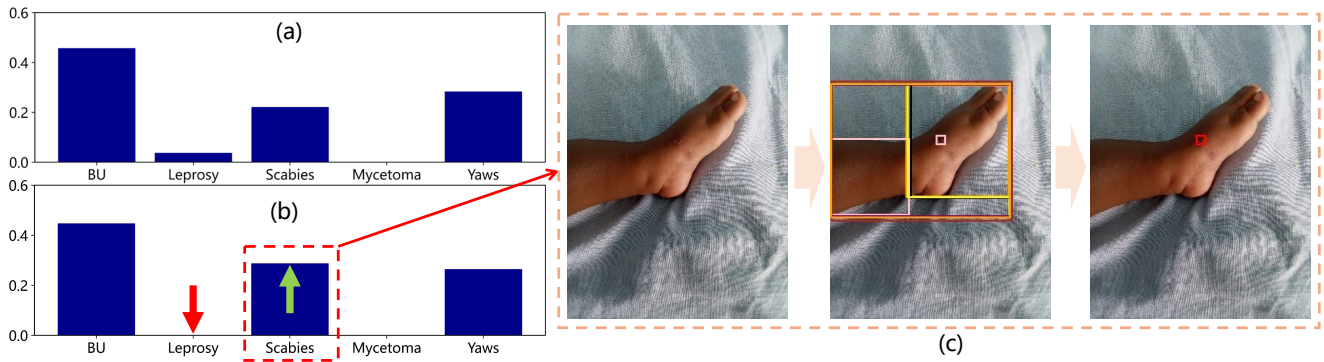
Figure 4. Performance comparison between baseline(global) and fusion(local + global) methods with the MIND-the-SKIN dataset. **(a)** shows the confidence for each condition by the baseline method; **(b)** the class-wise confidence by our fusion method; **(c)** demonstrates one "Scabies" example that is wrongly recognized by the baseline method and correctly predicted by our proposed fusion model.
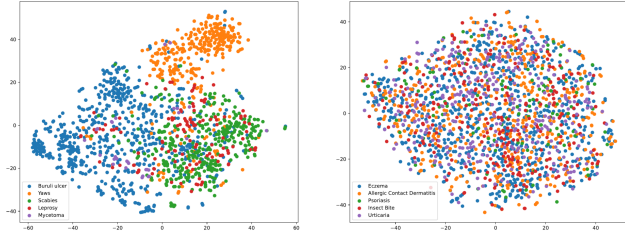

Figure 5. Visualization of data separability for the NTD (left) and SCIN (right) datasets after feature extraction by ViT.

sible explanation for the suboptimal performance is the difficulty in distinguishing between the conditions represented in the dataset, as shown by the t-SNE plots after feature extraction using ViT (Figure 5). Additionally, the dataset does not provide definitive labels for each image, but rather a list of potential labels with associated confidence scores, leading to label noise. Thus, while SCIN is the best available public dataset for our experiments, it remains challenging to use effectively, and its optimal usage is still unclear.

4.4. Interpretable Results

Figure 6 shows examples of "Buruli ulcer" and "Yaws," where the diseased areas are easily identifiable on the skin.

The distribution of the row scores in the CAM indicates that both cases have a high probability of the correct disease and low confidence in others. The column score distribution demonstrates that our model focuses more on the lesion part rather than the whole body. The selected bounding box (third column of the figure) and segmentation mask (fourth column) provide professionals with an intuitive yet reliable explanation for the prediction. Note that we report the final prediction using the top-5 visual concepts, and we select the top-1 from these to represent the visual concept that contributes most to the prediction, considering that multiple concepts might confuse the medical diagnosis.

4.5. Ablation Study

Table 4 indicates the performance comparison of different feature combinations on the MIND-the-SKIN dataset with train/total=0.5. The "Local-only" and "Baseline" denote the methods that only use local or global features individually for the final prediction. The "Concatenation-1" represents that we simply concatenate global feature to n local features as $n+1$ dimensional feature $\mathbf{O}' \in \mathbb{R}^{N \times (n+1) \times D}$. "Concatenation-2" follows the feature fusion method in [37] to combine local and global features in
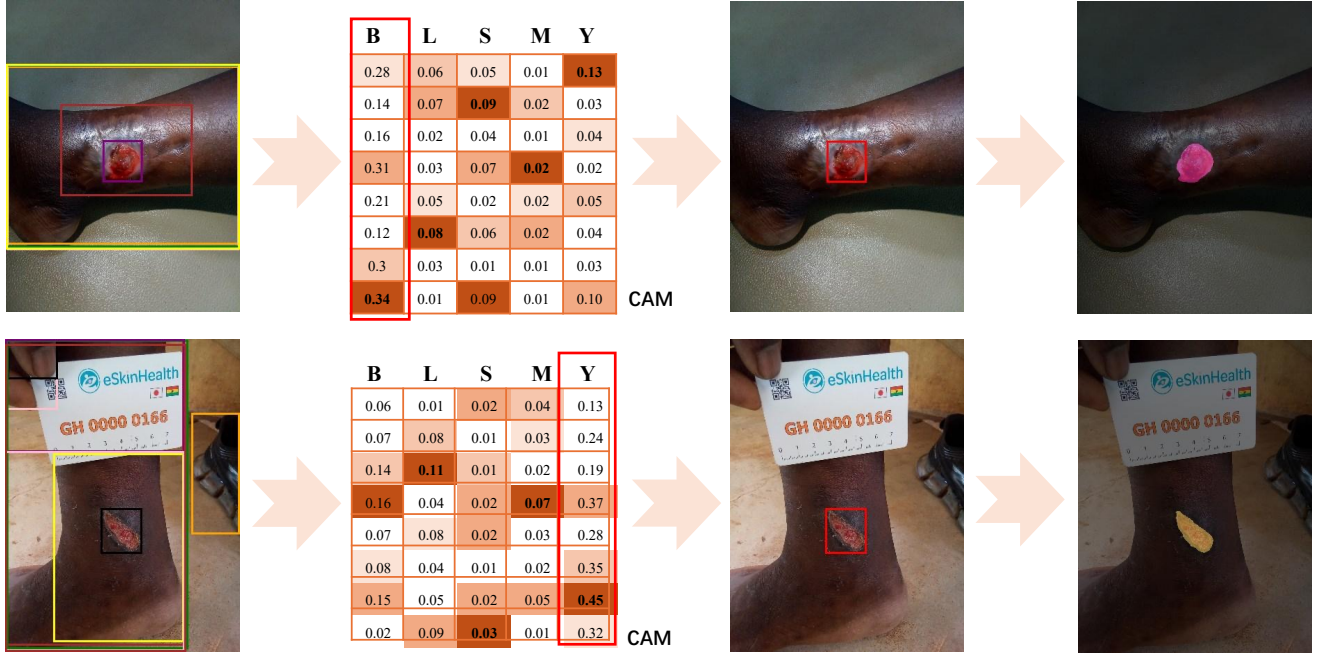
Figure 6. Qualitative results of skin disease diagnosis in the MIND-the-SKIN dataset. The first row represents one example from the "Buruli ulcer" disease; The second row demonstrates the example from "Yaws" disease. Both samples select the most contributed visual concepts and prove the effectiveness of our proposed method.

Table 4. Comparison of different feature combining methods on the MIND-the-SKIN dataset with train/total=0.5.

Methods	Precision	Recall	F1	Accuracy
Local-only	0.598	0.589	0.593	0.69
Baseline (global-only)	0.550	0.565	0.557	0.764
Concatenation-1	0.576	0.575	0.57	0.676
Concatenation-2	0.555	0.571	0.562	0.671
Average Sum	0.609	0.534	0.536	0.734
Ours	**0.651**	**0.631**	**0.641**	**0.809**

Table 5. Comparison of various k on the MIND-the-SKIN dataset.

k	Precision	Recall	F1	Accuracy
1	0.621	0.612	0.616	0.806
2	0.615	0.624	0.619	0.807
5	**0.651**	**0.631**	**0.640**	**0.809**
6	0.647	0.634	0.640	0.808
15	0.607	0.591	0.599	0.804

5. Conclusion

In this paper, we leverage the foundation AI model, SAM, to automatically segment visual skin images and propose a cross-attention model designed to harness complementary information between local visual concepts and global features in challenging clinical skin disease images. To effectively explain our model's decision-making process, we integrate CAM and multi-instance learning to identify the most influential concepts, which are generated by SAM using stochastic text prompts. Our experiments demonstrate that the proposed method consistently outperforms competitive approaches across various metrics, underscoring the effectiveness of the dual-branch design. In addition, our method provides better interpretability, offering explainable predictions that enhance the reliability of AI-based diagnoses for medical professionals. This interpretability is crucial for building trust and improving diagnostic accessibility in future applications.

feature space as $\mathbf{O}'' \in \mathbb{R}^{N \times n \times 2D}$. "Average Sum" conducts average pooling and sum operation on local and global features. The comparison results demonstrate that our cross-attention fusion method achieves better performance than other fusion strategies.

Table 5 presents a performance comparison using different top-k strategies. It is observed that the model achieves optimal performance when $k = 5$. While fewer visual concepts are selected, there is a significant drop in precision, recall, and F1 score, even though the accuracy does not vary much. This can be attributed to the fact that accuracy is the metric used to select the best model, and the limited samples from the training dataset can lead to overfitting on this metric. Furthermore, when $k = 15$, which is half the total number of concepts, there is a noticeable decline in all performance metrics. This decline occurs because including more visual concepts incorporates background noise from SAM into the final prediction, degrading performance.

References

[1] Mind-the-skin project: Multi-functional innovative digital toolkit for the skin diseases in lmics and beyond. 5

[2] Segment and track anything. https://github.com/z-x-yang/Segment-and-Track-Anything. 3

[3] Titus Brinker, Achim Hekler, Alexander Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Jochen Utikal, Christof Kalle, Stefan Fröhling, and Michael Weichenthal. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 03 2019. 1

[4] M Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: overview and future directions. *IEEE journal of biomedical and health informatics*, 23(2):474–478, 2019. 1

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 2

[6] Alberto Coustasse, Raghav Sarkar, Bukola Abodunde, Brandon J. Metzger, and Chelsea M. Slater. Use of teledermatology to improve dermatological access in rural areas. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, 2019. 1

[7] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5, 6

[9] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. 1

[10] Karoline Freeman, Jacqueline Dinnes, Naomi Chuchu, Yemisi Takwoingi, Sue E Bayliss, Rubeta N Matin, Abhilash Jain, Fiona M Walter, Hywel C Williams, and Jonathan J Deeks. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *bmj*, 368, 2020. 1

[11] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585, 2015. 5

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021. 5

[14] Md Kamrul Hasan, Lavsen Dahal, Prasad N Samarakoon, Fakrul Islam Tushar, and Robert Martí. Dsnet: Automatic dermoscopic skin lesion segmentation. *Computers in biology and medicine*, 120:103738, 2020. 2

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[18] Galib Muhammad Shahriar Himel, Md Masudul Islam, Kh Abdullah Al-Aff, Shams Ibne Karim, and Md Kabir Uddin Sikder. Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system. *International Journal of Biomedical Imaging*, 2024(1):3022192, 2024. 2, 3

[19] Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. Skinsam: Empowering skin cancer segmentation with segment anything model. *arXiv preprint arXiv:2304.13973*, 2023. 2, 3

[20] Xin Hu, Kai Li, Deep Patel, Erik Kruus, Martin Renqiang Min, and Zhengming Ding. Weakly-supervised temporal action localization with multi-modal plateau transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2024. 4

[21] Shubham Innani, Prasad Dutande, Ujjwal Baid, Venu Pokuri, Spyridon Bakas, Sanjay Talbar, Bhakti Baheti, and Sharath Chandra Guntuku. Generative adversarial networks based skin lesion segmentation. *Scientific Reports*, 13(1):13467, 2023. 2

[22] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1, 2

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[24] Baiying Lei, Zaimin Xia, Feng Jiang, Xudong Jiang, Zongyuan Ge, Yanwu Xu, Jing Qin, Siping Chen, Tianfu Wang, and Shuqiang Wang. Skin lesion segmentation via generative adversarial networks with dual discriminators. *Medical Image Analysis*, 64:101716, 2020. 2

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[26] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S Corrado, Lily H Peng, Dale R Webster, Dennis Ai, Susan J Huang, Yun Liu, R Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, June 2020. 1

[27] Priya Narayanan, Xin Hu, Zhenyu Wu, Matthew D Thielke, John G Rogers, Andre V Harrison, John A D'Agostino, James D Brown, Long P Quang, James R Uplinger, et al. A multi-purpose realistic haze benchmark with quantifiable haze levels and ground truth. *IEEE Transactions on Image Processing*, 32:3481–3492, 2023. 2

[28] Şaban Öztürk and Umut Özkaya. Skin lesion segmentation with improved convolutional neural network. *Journal of digital imaging*, 33:958–970, 2020. 2

[29] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020. 5

[30] Luis R Soenksen, Timothy Kassis, Susan T Conover, Berta Marti-Fuster, Judith S Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R Stavert, Caroline C Kim, Maryanne M Senna, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581):eabb3652, 2021. 1

[31] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, 8, 1995. 2

[32] Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. 2

[33] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*, 26(8):1229–1234, 2020. 1, 2

[34] Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2024. 1

[35] Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep Kumar S, Tiya Tiyasirichokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Crowdsourcing dermatology images with google search ads: Creating a real-world skin condition dataset, 2024. 5

[36] Fengying Xie, Jiawen Yang, Jie Liu, Zhiguo Jiang, Yushan Zheng, and Yukun Wang. Skin lesion segmentation using high-resolution convolutional neural network. *Computer methods and programs in biomedicine*, 186:105241, 2020. 2

[37] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020. 7

[38] Jufeng Yang, Xiaoping Wu, Jie Liang, Xiaoxiao Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Self-paced balance learning for clinical skin disease recognition. *IEEE transactions on neural networks and learning systems*, 31(8):2832–2846, 2019. 5

[39] Kashan Zafar, Syed Omer Gilani, Asim Waris, Ali Ahmed, Mohsin Jamil, Muhammad Nasir Khan, and Amer Sohail Kashif. Skin lesion segmentation from dermoscopic images using convolutional neural network. *Sensors*, 20(6):1601, 2020. 2

[40] Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23385–23395, 2024. 3