SANO: Score-based Diffusion Model for Anomaly Localization in Dermatology

Alvaro Gonzalez-Jimenez [1] Simone Lionetti [2] Marc Pouly [2]
Alexander A. Navarini [1,3]
University of Basel [1], Lucerne University of Applied Sciences and Arts [2],
University Hospital of Basel [3]
alvaro.gonzalezjimenez@unibas.ch, {simone.lionetti, marc.pouly}@hslu.ch,
alexander.navarini@usb.ch

Abstract

Supervised learning for dermatology requires a large volume of annotated images, but collecting clinical data is costly, and it is virtually impossible to cover all clinical cases. Unsupervised anomaly localization circumvents this problem by learning the healthy data distribution. However, algorithms which use a generative model and localize pathologic regions based on a reconstruction error are not robust to domain shift, which is a problem for dermatology due to the low level of standardization expected in many applications. Our method, SANO, uses score-based diffusion models to produce a log-likelihood gradient map highlighting areas that contain abnormalities. A segmentation mask can then be calculated based on deviations from typical values observed during training. After benchmarking SANO on an industrial dataset, we train it on a public non-clinical dataset of healthy hand images without ornaments, evaluate it on the task of detecting jewelry within images from the same dataset, and prove its robustness by using it on clinical pictures to localize hand eczema. We demonstrate that SANO outperforms competing approaches from the literature without introducing additional computational costs.

1. Introduction

Skin diseases are a major concern globally, accounting for a large number of clinic visits. In developing countries, the lack of experts to diagnose and treat these conditions is a critical issue, with a ratio of dermatologists to the general population which can be as low as 1 to 216,000 [12]. This led to significant interest in developing systems capable of identifying and diagnosing skin diseases, also involving large organizations, with most efforts relying on supervised Deep Learning (DL) algorithms [23].

However, the requirement for large amounts of annotated data poses significant challenges in the field of dermatology. Although simpler than in other medical fields, the collection of images is affected by uncontrolled acquisition conditions such as camera model, lighting, and view angle. Currently, no established method exists to standardize images collected under these varied conditions. Furthermore, most of the current training data consist of white skin samples, leading to a significant bias in the performance of DL algorithms, particularly when applied to different skin tones [1, 15, 18]. This is a major obstacle to the deployment of teledermatology in emerging countries and raises questions about fairness for ethnic minorities. The acquisition of sufficient data for rare pathologies is also a challenge, particularly given the strong geographical dependence on the data distribution. For example, insect bites are common in Africa but rare in Europe [21, 33]. In addition, annotation is a time-consuming task that requires the expertise of clinical professionals, making the process costly, particularly for obtaining detailed segmentation masks. Finally, the annotation process introduces human bias, as demonstrated by low inter-annotator agreement in the field [27], whereas the gold standard for dermatologic diagnosis is often histopathology, which raises ethical concerns when a biopsy is not clinically necessary.

Learning the appearance of healthy skin to locate abnormal regions is an approach that alleviates many of the difficulties listed above. This approach, called *unsupervised anomaly localization*, is sometimes referred to as semi-supervised if the training data is filtered to be free of unhealthy examples. Despite its potential in dermatology, where images of healthy skin are easily available, this approach has limitations. It cannot produce a diagnosis and often leads to less accurate segmentation masks.

To locate anomalies, typical unsupervised approaches use a generative model to reconstruct healthy images. The

difference between an image and its reconstructed version is then used to identify lesions. Researchers have explored the combination of unsupervised anomaly localization with various generative models such as Variational Autoencoders (VAEs) [4, 6, 8, 36], Generative Adversarial Networks (GANs) [3, 5, 28], and Diffusion Models [24, 32, 34]. While these methods perform well in highly standardized medical imaging settings, they struggle in less controlled conditions [16]. Recently, alternative strategies for unsupervised anomaly localization have been proposed. These include works that investigate the use of gradients of the log likelihood with respect to inputs from EBMs to create a normalcy score heatmap [13, 38], and other methods that explore patch-based approaches to anomaly localization [10, 35].

In this work, we present Score-based ANOmaly localization (SANO), a new method for unsupervised anomaly localization that leverages score-based diffusion models. These models have been shown to achieve state-of-the-art likelihood values by directly approximating the gradients of the log likelihood [30]. Our approach is unique in combining the idea of using log-likelihood gradients for anomaly localization with score-based diffusion models which are trained to estimate precisely these gradients. Notably, SANO does not require reconstruction, whose computational complexity is one of the main drawbacks of score-based diffusion models. To the best of our knowledge, and including recent reviews [9], this is the first work that achieves unsupervised anomaly localization by combining the two above-mentioned ideas, noting that they are particularly suited to be applied together.

In the absence of a standard for medical anomaly detection, we first evaluate SANO on the MVTec benchmark [7]. We then apply SANO to localize jewelry on healthy hands in the 11k Hands dataset [2, 14]. This scenario, although not a clinical task, is potentially relevant for digital dermatology workflows where the presence of extraneous objects can reduce suitability of data or violate data anononymization policies. Finally, and most importantly, we consider anomaly localization for the segmentation of hand eczema in a private clinical dataset without retraining, under fairly standard but different conditions from those of 11k Hands. We demonstrate that SANO is significantly more robust than all other considered methods under this domain shift, making it a promising candidate for disease-agnostic segmentation of pathological skin for digital dermatology.

2. Methods

2.1. Score-based diffusion models

Several generative modeling approaches were recently unified under a single framework and grouped under the common name of *score-based diffusion models* [30]. Models which belong to this class are associated with a stochastic process $\mathbf{x}(t)$ indexed by a time variable $t \in [0, 1]$ which progressively maps a data point $\mathbf{x}(0)$ to a sample $\mathbf{x}(1)$ from a prior distribution $p_1(\mathbf{x})$ representing random noise. The transformation of data into noise admits a reverse process that enables mapping a sample $\mathbf{x}(1)$ from the prior to a data point $\mathbf{x}(0)$ following the data distribution $p_0(\mathbf{x})$, *i.e.* it constitutes a generative model. This reversible transformation process from $\mathbf{x}(0)$ to $\mathbf{x}(1)$ is defined by a Stochastic Differential Equation (SDE) and induces a one-parameter family of probability distributions $p_t(\mathbf{x})$. The family smoothly interpolates between the $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$ and its evolution with respect to t may be factorized into the product with a transition kernel $p_{t'}(\mathbf{x}') = p_{tt'}(\mathbf{x}'|\mathbf{x})p_t(\mathbf{x})$.

The training process for score-based diffusion models consists in finding an approximation $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ for the gradient of the log likelihood with respect to the inputs, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, which is also called the (*Stein*) score function [22, 31] of $p_t(\mathbf{x})$. The space of approximating functions $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ parametrized by $\boldsymbol{\theta}$ is often taken to be a given deep neural network architecture. The matching can be achieved by minimizing the loss

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2}\int_0^1 \mathbb{E}_{p_{0t}(\mathbf{x}'|\mathbf{x})p_0(\mathbf{x})} \left[\|\nabla_{\mathbf{x}'} \log p_{0t}(\mathbf{x}'|\mathbf{x}) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', t)\|_2^2\right] \mathrm{d}t. \quad (1)$$

Note that, in this formulation, the analytic or numeric tractability of the normalization factor for the time-dependent probability distribution $p_t(\mathbf{x})$ is irrelevant. Remarkably, it has been shown that although score-based diffusion models do not directly optimize the likelihood of the data, there is a way of weighting the integrand in (1) which turns $\mathcal{J}(\boldsymbol{\theta})$ into a lower bound for the likelihood [29, 30]. Empirical results in the same references demonstrate that score-based diffusion models obtain very competitive likelihood values on a range of practical tasks.

The cited works on score-based diffusion models considered three types of SDEs: Variance Exploding (VE), Variance Preserving (VP), and sub-VP. This work will consider the VP SDE, the simplest apart from VE which empirically delivers worse likelihoods. The equation reads

$$\mathrm{d}\mathbf{x}(t) = -\frac{1}{2}\beta(t)\mathbf{x}(t)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}(t), \quad (2)$$

where $\mathbf{w}$ denotes the standard Wiener process and $\beta(t)$ is a positive function. Following [17, 30], we also set

$$\beta(t) = \bar{\beta}_{min} + t(\bar{\beta}_{\max} - \bar{\beta}_{\min}). \quad (3)$$

In particular, we note that these definitions yield a gaussian transition kernel, which significantly simplifies calculations and indicates that stochastic process evolution from $t = 0$ to $t = 1$ corresponds to gradual addition of gaussian noise.

2.2. Anomaly localization with scores

The method for locating anomalies in images we use is based on the idea that gradients of the log likelihood with respect to input values are typically larger for inputs that are unlike any training examples. Our approach is built on score-based diffusion models, which are trained to approximate these gradients directly for all values of t, and can be used to estimate them for data samples by simply evaluating the approximated function $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ at $t = 0$. This corresponds to a single forward pass, in contrast to the Energy-Based Models (EBMs) used in [13] which require backpropagation during inference.

To obtain anomaly scores, we first combine the color channels into the square of the gradient vector for each pixel. We then apply a Gaussian filter to the gradient heatmap to increase the scale that defines an anomaly while retaining pixel-level resolution. The filter scale can be tuned using a validation set, but in our approach we fix $\sigma = 4$ based on an initial guess of the minimal resolution scale for anomaly masks, which are unlikely to be determined by isolated pixels.[1] Our approach models the empirical gradient distribution for normal data as a zero-centered Gaussian, isotropic in color, with the same variance for all pixels. Unlike [13], we did not find any benefit in normalizing gradients pixel by pixel.

Once the anomaly score heatmap has been obtained, generating a segmentation map involves setting a threshold. Any density estimation technique applied to the distribution of scores of a validation set can in principle be used for this task. In cases where a validation set containing both normal and anomalous data is available, the appropriate objective can be optimized to determine the threshold. If only normal data is available, the threshold may be chosen by setting the expected false positive rate among the validation examples.

Alternatively, if a validation set is not available, a heuristic recipe can be used, assuming the score distribution is centered around zero. Anomalies can be identified as those pixels whose gradients deviate from zero by more than a certain number of standard deviations. This recipe provides a reasonably conservative normalcy criterion, even for non-Gaussian distributions, as the Mahalanobis squared norm is dominated by the longest tails.

3. Experiments

In this section, we demonstrate the capacity of SANO to localize anomalies within images without supervision. To this end, first we evaluate SANO with competing approaches on an industrial benchmark dataset. Then we repeat the comparison for finding jewelry on images of healthy hands, and for segmenting dermatologic lesions on a different dataset without retraining.

3.1. Datasets

MVTec [7] is a benchmark dataset for anomaly localization in the context of industrial inspection. It consists of 15 different object and texture categories and is widely used to evaluate the performance of anomaly localization algorithms. The dataset provides pre-defined splits for training and testing. Since this dataset is usually evaluated with metrics which do not require a threshold, we do not require a validation set. To asses the performance of our method, we train and evaluate on a separate model for each texture and object.

11k Hands [2, 14] is a public dataset that contains 11,076 hand images from 190 subjects with a resolution of 1600×1200 pixels. Each hand was photographed from the dorsal and palmar sides with a uniform white background and the same indoor lighting, approximately at the same distance from the camera. We train models on 5,589 hand images without jewelry, use a validation set of 1,022 images (324 without and 698 with jewels) to determine the threshold and a final test set of 4,434 images (1953 without and 2481 with jewels) to evaluate the localization of jewelry. The ratio jewels pixels is 0.49% in the test set. Finally, the splits were always grouped by subjects with fixed ratios to avoid data leakage.

PhotoBox is a private dataset collected at a university hospital[2] containing images of hand eczema from a total of 131 patients. Both hands were photographed together from the dorsal and palmar sides for each patient. The pictures were taken in a closed structure where hands are inserted through a slit, illuminated by LEDs, and the background is a uniform green surface. The camera was connected to a tablet to guide the patient in the placement of the hands. All images, which have a resolution of 3456×2304, were manually annotated by an expert dermatologist. We split the sets of images with anomalies into two parts of equal size and used one for tuning the threshold parameters and the other for evaluation. As in the previous case, the splits were always grouped by patient images.

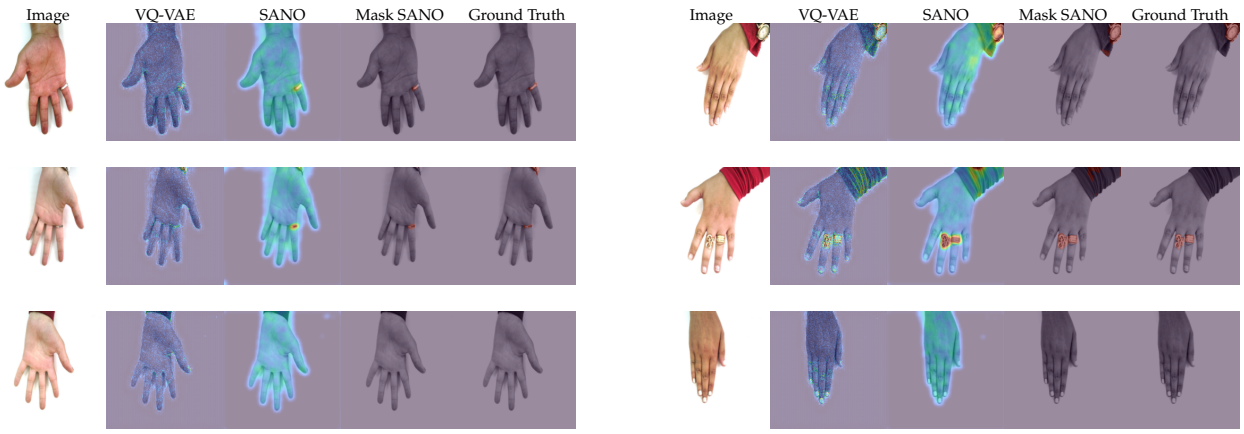
3.2. Training

To train the score-based models, we resize all images to 256×256 pixels and do not employ data augmentation. We approximate the score using a U-Net for $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ as suggested by [11]. We set the training objective as in Eq. (1) with the choice in Eq. (3), 1000 diffusion time steps, $\beta_{\text{min}} = 10^{-4}$, and $\beta_{\text{max}} = 0.02$, using a public codebase[3]

[1] To ensure a fair comparison, for all hand models we choose whether to apply this smoothing based on a validation set.

[2] To be replaced with the explicit name after double-blind review.

[3] https://github.com/yang-song/score_sde_pytorch



(a) Palmar (b) Dorsal

Figure 1. Heatmap of anomalies in the 11k Hands dataset, comparing SANO and the second-best model (VQ-VAE). The figure shows the original image, anomaly detections by both models, the binarize mask of SANO, and the ground truth segmentation mask. Warmer colors indicate higher anomaly scores. The ground truth segmentation mask provides a reference for the accuracy of the model's anomaly detections.

adapted for the purpose. We set the batch size to 64 and optimize for 400k iterations with the Adam [19] optimizer and a learning rate of 2×10^{-4}.

Alongside the score-based diffusion model we train several other Deep Learning (DL) models for comparison. More specifically, we consider a simple Autoencoder (AE), a Variational Autoencoder (VAE) [20, 26], and a Context-encoding Variational Autoencoder (ceVAE) [37] taken from the repository for the Medical Out-Of-Distribution Challenge of MICCAI 2020;[4] VQ-VAE from [25]; and AnoVAEGAN, which was proposed in [6].

3.3. Evaluation

We first evaluate the performance of anomaly localization models for finding defects in the industrial images of MVTec and jewelry on the healthy hands of 11k Hands. In these two cases, we use the same preprocessing pipeline as for training. Then, to investigate the robustness with respect to domain shifts for the models trained on 11k Hands, we use them to detect hand eczema in the PhotoBox dataset. In order to keep domain shift sizeable and at the same time to have a chance of success, we purposefully ignore that in PhotoBox both hands are captured simultaneously (in contrast to 11k Hands), but we use color segmentation to change the green background into uniform white (as is the case in 11k Hands).

We evaluate the performance of the different anomaly scoring systems using standard metrics which do not require a threshold, namely the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic curve (AUROC).

[4] https://github.com/MIC-DKFZ/mood

For obtaining segmentation metrics, we selected a threshold that maximized the F1 score in the corresponding validation set. This threshold value was then used to generate binary masks for each image in the test set, which allowed us to calculate the best Dice coefficient and the best Intersection over Union (IoU).

4. Results

4.1. Segmentation of anomalies in MVTec

First, we benchmark SANO for anomaly localization on MVTec. As shown in Tab. 1, most algorithms achieve a good result in this setting. In particular, SANO outperforms EBM [13], even though both use the gradients of the log likelihood to obtain the affected area. This suggests that modelling the score $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ directly could be an advantage in terms of metrics besides reducing computational cost. For textures, SANO obtains better results compared to the considered baselines, and in the case of objects, it maintains good performance except on the cable and transistor images.

4.2. Segmentation of jewelry

In the broad context of dermatology, we present results for anomaly localization applied to finding jewelry in the 11k Hands dataset. All considered algorithms achieve reasonably good performance, as shown in Tab. 2. The reported uncertainties are estimates of expected variations due to the finite size of the evaluation set, computed as the standard deviations of 1000 bootstrap runs where random selection with replacement was stratified over individuals in the dataset. We observe that SANO outperforms all considered

	Category	SSIM-AE	l_2-AE	AnoGAN	EBM	SANO
Texture	Carpet	**0.87**	0.59	0.54	0.63	0.84
	Grid	0.94	0.90	0.58	0.86	**0.97**
	Leather	0.78	0.75	0.64	0.87	**0.99**
	Tile	0.59	0.51	0.50	0.57	**0.91**
	Wood	0.73	0.73	0.62	0.74	**0.86**
Object	Bottle	**0.93**	0.86	0.86	0.72	0.81
	Cable	0.82	**0.86**	0.78	0.56	0.55
	Capsule	**0.94**	0.88	0.84	0.64	0.76
	Hazelnut	0.97	0.95	0.87	0.78	**0.98**
	Metal Nut	**0.89**	0.86	0.76	0.65	0.66
	Pill	0.91	0.85	0.87	0.75	**0.98**
	Screw	**0.96**	**0.96**	0.80	0.87	0.95
	Toothbrush	0.92	**0.93**	0.90	0.68	0.84
	Transistor	**0.90**	0.86	0.80	0.74	0.61
	Zipper	0.88	0.77	0.78	0.55	**0.92**

Table 1. AUROC results for anomaly localization on MVTec. A model was trained for each texture/object separately. Results from other benchmarks taken from [7, 13].

Model	AUROC	AUPRC	Dice	IoU
AE	0.946(1)	0.123(14)	0.231(14)	0.130(9)
VAE	0.937(2)	0.131(11)	0.227(11)	0.149(7)
ceVAE	0.941(2)	0.102(16)	0.173(16)	0.095(9)
VQ-VAE	0.946(2)	0.416(22)	0.448(15)	0.289(10)
AnoVAEGAN	0.943(2)	0.101(12)	0.189(15)	0.091(9)
SANO	**0.963(2)**	**0.422(10)**	**0.551(13)**	**0.383(8)**

Table 2. Scores for unsupervised jewelry localization on the 11k Hands dataset. The standard deviation over 1000 bootstrap runs is reported in brackets as the uncertainty on the last digits. The best results are highlighted in **bold**.

reconstruction-based approaches. Some example masks obtained with SANO are illustrated in Fig. 1. The model is able to correctly segment jewels on both the dorsal and palmar sides of hands, and gets qualitatively worse results on wrist jewelry. Note that sleeves without jewels are not considered anomalies as they are present in the training set.

4.3. Segmentation of hand eczema

Finally, in Tab. 3 we report the results of the models trained on jewelry-free, healthy hands of 11k Hands on the localization of hand eczema in the PhotoBox dataset, again including uncertainties from 1000 bootstrap runs.

As discussed in Sec. 3.3, switching from jewelry localization in 11k Hands to hand eczema segmentation in the PhotoBox dataset constitutes a sizeable domain shift. Reconstruction-based anomaly localization methods are deeply affected by the context change, as reflected by the considerable drop in scores and by the masks in Fig. 2 which highlight small shifts in the reconstruction. The situation is significantly better for SANO which, despite a noticeable worsening of performance, still obtains good results and achieves a Dice score of 0.358 and an IoU of 0.231. Looking at Fig. 2 we can indeed see that SANO correctly marks the region where the lesion is located.

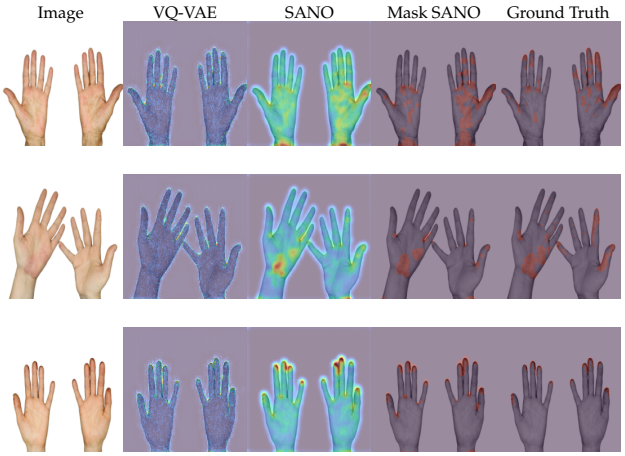


Figure 2. Heatmap of anomalies in the PhotoBox dataset, comparing SANO and the second-best model (VQ-VAE). The figure shows the original image, anomaly detections by both models, the binarize mask of SANO, and the ground truth segmentation mask. Warmer colors indicate higher anomaly scores. The ground truth segmentation mask provides a reference for the accuracy of the model's anomaly detections.

Model	AUROC	AUPRC	Dice	IoU
AE	0.641(3)	0.089(10)	0.151(9)	0.081(8)
VAE	0.634(5)	0.094(7)	0.153(13)	0.085(4)
ceVAE	0.714(5)	0.103(8)	0.178(11)	0.094(10)
VQ-VAE	0.819(4)	0.151(11)	0.280(14)	0.140(9)
AnoVAEGAN	0.769(3)	0.119(4)	0.184(3)	0.099(12)
SANO	**0.912(3)**	**0.268(11)**	**0.358(12)**	**0.231(10)**

Table 3. Scores for unsupervised eczema segmentation on the Photobox dataset. The standard deviation over 1000 bootstrap runs is reported in brackets as the uncertainty on the last digits. The best results are highlighted in **bold**.

5. Conclusions

This research paper introduced SANO, a novel method for unsupervised anomaly localization that utilizes the log-likelihood gradient magnitude from score-based diffusion models. Unlike other approaches based on generative modeling, SANO does not rely on reconstruction to identify anomalous regions.

After demonstrating that SANO is competitive on a public benchmark dataset of industrial defects, our study focused on learning the characteristics of healthy hands without jewelry from the public 11k Hands dataset. These were

then used to predict anomalous regions on photographs with and without jewelry from the same dataset, and on clinical images of hands affected by eczema from a somewhat similar but different context. The results show that SANO outperforms several other unsupervised anomaly localization methods in the same-domain images and its performance is superior by a large margin in case of a domain shift.

These observations demonstrate that SANO is an important step in developing DL solutions for digital dermatology which work under a wide range of conditions both for clinical tasks such as skin lesion segmentation and non-clinical goals such as guaranteeing image quality and preserving patient privacy

References

[1] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, 154(11):1247, 2018. 1

[2] Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019. 2, 3

[3] Simon Andermatt, Antal Horváth, Simon Pezold, and Philippe Cattin. Pathology Segmentation Using Distributional Differences to Images of Healthy Origin. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 228–238. Springer International Publishing, 2019. 2

[4] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, page 101952, 2021. 2

[5] Christoph Baur, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 718–727. Springer International Publishing, 2020.

[6] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 161–169. Springer International Publishing, 2019. 2, 4

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. 2, 3, 5

[8] Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using Constrained Adversarial Auto-encoders. In *MIDL Conference Book*. MIDL, 2018. 2

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. 2022. 2

[10] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[12] Ncoza Dlova, A Chateau, N Khoza, A Skenjane, Mondli Mkhize, O Katibi, A Grobler, Joyce Tsoka-Gwegweni, and A Mosam. Prevalence of skin diseases treated at public referral hospitals in KwaZulu-Natal, South Africa. *The British journal of dermatology*, 178, 2017. 1

[13] Ergin Utku Genc, Nilesh Ahuja, Ibrahima J Ndiour, and Omesh Tickoo. Energy-based anomaly detection and localization. *arXiv preprint arXiv:2105.03270*, 2021. 2, 3, 4, 5

[14] Alvaro Gonzalez-Jimenez, Simone Lionetti, Ludovic Amruthalingamnd, Philippe Gottfrois, Marc Pouly, and Alexander Navarini. Jewelry segmentation masks for the 11k Hands dataset. 2022. 2, 3

[15] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828. IEEE.

[16] Matthäus Heer, Janis Postels, Xiaoran Chen, Ender Konukoglu, and Shadi Albarqouni. The OOD blind spot of unsupervised anomaly detection. In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 286–300. PMLR, 07–09 Jul 2021. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2

[18] Louis Henry Kamulegeya, Mark Okello, John Mark Bwanika, Davis Musinguzi, William Lubega, Davis Rusoke, Faith Nassiwa, and Alexander Börve. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. 2019. 1

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017. 4

[20] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. 2014. 4

[21] Samson K. Kiprono, Julia W. Muchunu, and John E. Masenga. Skin diseases in pediatric patients attending a tertiary dermatology hospital in Northern Tanzania: A cross-sectional study. *BMC Dermatology*, 15(1):16, 2015. 1

[22] Qiang Liu, Jason Lee, and Michael Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 276–284. PMLR, 2016. 2

[23] 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, 2020. 1

[24] Walter H. L. Pinaya, Mark S. Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models, 2022.

[25] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. 2019. 4

[26] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. 4

[27] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. *Handling Inter-Annotator Agreement for Automated Skin Lesion Segmentation*. 2019. 1

[28] Thomas Schlegl, Philipp Seeböck, Sebastian Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. F-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks. *Medical Image Analysis*, 54, 2019. 2

[29] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, 2021. 2

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR*, 2021. 2

[31] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *The Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6.2, pages 583–603. University of California Press, 1972. 2

[32] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion Models for Medical Anomaly Detection. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 35–45. Springer Nature Switzerland, 2022. 2

[33] World Health Organization. Epidemiology and management of common skin diseases in children in developing countries. (WHO/FCH/CAH/05.12), 2005. 1

[34] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 650–656, June 2022. 2

[35] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[36] Suhang You, Kerem C. Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 540–556. PMLR, 08–10 Jul 2019. 2

[37] David Zimmerer, Simon A. A. Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. 2018. 4

[38] David Zimmerer, Jens Petersen, Simon A. A. Kohl, and Klaus H. Maier-Hein. A case for the score: Identifying image anomalies using variational autoencoder gradients, 2019. 2