2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761412

OUT-OF-DISTRIBUTION DETECTION IN DERMATOLOGY USING INPUT PERTURBATION AND SUBSET SCANNING

Hannah Kim *Girmaw Abebe Tadesse* *Celia Cintas* *Skyler Speakman* *Kush Varshney*

IB Research Africa
Kenya Lab

ABSTRACT

Recent advances in deep learning have led to breakthroughs in the development of automated skin disease classification. As we observe an increasing interest in these models in the dermatology space, it is crucial to address aspects such as the robustness towards input data distribution shifts. Current models tend to make incorrect inferences for test samples from different hardware devices and clinical settings or unknown disease samples, which are out-of-distribution (OOD) from the training samples. To this end, we propose a simple yet effective approach that detects these OOD samples prior to making any decision. The detection is performed via scanning in the latent space representation (e.g., activations of the inner layers of any pre-trained skin disease classifier). The input samples are also perturbed to maximise divergence of OOD samples. We validate our OOD detection approach in two use cases: 1) identify samples collected from different protocols, and 2) detect samples from unknown disease classes. Our experiments yield competitive performance across multiple datasets for both use cases. As most skin datasets are reported to suffer from bias in skin tone distribution, we further evaluate the fairness of these OOD detectors across different skin tones.

***Index Terms*—** Subset scanning, Skin disease classification, Out-of-distribution sample detection

1. INTRODUCTION

Skin disease remains a global health challenge, with skin cancer being the most common cancer worldwide [1]. Following the recent success of deep learning (DL) in various computer vision problems, convolutional neural networks (CNNs) have been employed for skin disease classification. As we observe increasing interest in DL for dermatology [2, 3], it is imperative to address transparency, robustness, and fairness of these solutions [4, 5, 6]. While many existing DL techniques [7, 8] achieve high performance on publicly available datasets [1, **?**, 9, 10], they utilize ensembles of multiple models aimed at maximising performance with limited consideration to shifts in the input data [8, 11, 7]. This might result in incorrectly classifying new samples with high confidence though these samples might be from previously unknown classes. Thus, it is necessary to detect out-of-distribution (OOD) samples prior to making decisions in order to achieve principled transfer of knowledge from in-distribution (ID) training samples to OOD test samples, thereby extending the usability of the models to previously unseen scenarios.

We propose a simple yet effective approach that scans over the activations of the inner layers of any pre-trained skin disease classifier to detect OOD samples. Input data is perturbed beforehand with our proposed ODIN$_{low}$, a modification of ODIN [12], which improve the OOD detection performance in earlier layers of the network. In our framework, we define two different OOD use cases: *protocol variations* (e.g., different hardware devices, lighting settings and not compliant with clinical protocol); and *unknown disease types* (e.g., samples from new disease type that was not observed during training). Without requiring any prior knowledge of the OOD samples, our proposed approach performs comparably or better compared to the existing OOD detectors, Softmax Score [13] and ODIN [12], for both use cases. We further explored how our proposed and existing OOD detectors perform across skin tones to evaluate fairness. We show that the current OOD detectors show higher performance in detecting darker skin tones as OOD samples than those of lighter skin tones, which is likely impacted by the imbalanced training datasets that heavily lack samples of dark skin tones.

Table 1. OOD sample detectors for skin disease classification.

	[8]	[11]	[7]	[14]	[15]	[16]	[17]	[18]	ours
Ensemble	✓	✓	✓	✗	✗	✓	✗	✗	✗
Post-Training Detection	✗	✗	✗	✗	✓	✓	✓	✓	✓
New Protocol Detection	✗	✗	✗	✗	✗	✗	✓	✓	✓
New Disease Detection	✓	✓	✓	✓	✓	✓	✓	✗	✓
Algorithmic Fairness	✗	✗	✗	✗	✗	✗	✗	✗	✓

2. RELATED WORK

Our review of existing OOD detection methods is grouped into *pre-training* [8, 11, 7, 14] and *post-training* [15, 16, 17], based on where the detection step is applied. **Pre-training detection** approaches have prior knowledge of the OOD samples and incorporate it during their training phases. Many of these approaches utilize ensembles of existing CNNs (and their variants) to detect OOD samples [8, 11, 7, 14]. Ahmed *et al.* [8] and Bagchi *et al.* [14] applied one-class learning where each class was iteratively discarded as an OOD class in a one-vs-all cross-validation strategy, and the OOD samples were detected by taking the prediction average of all the models. Ensemble employed by Zhang *et al.* [7] consisted of both multi-class and binary classifiers for OOD detection. Gessert *et al.* [11] utilized an extra skin dataset of OOD samples to train an ensemble of CNNs. **Post-training detection** approaches do not require any prior knowledge of the OOD samples during training [15, 16, 17, 18]. Pacheco *et al.* [16] detected OOD samples using *Shannon entropy* [19] and *cosine similarity* metrics on their CNN's probability outputs. Combalia *et al.* [15] detected OOD samples using *Monte-Carlo Dropout* [20] and test data augmentationto estimate uncertainty such as entropy and variance in their predictions. Pacheco *et al.* [17] extended Gram-OOD [21] with layer-specific normalization of Gram Matrices to detect OOD samples. Zaida *et al.* [18] simulate OOD samples from in-distribution (ID) samples for training, and use K-reciprocal nearest neighbour during inference for OOD detection.

Table 1 summarizes notable OOD detectors in dermatology. Most detectors employed pre-training approaches, which require prior OOD knowledge, and used ensembles of CNNs, which could easily result in model complexity. We propose a simple, post-training detector that can be applied to any pre-trained network.

3. PROPOSED FRAMEWORK

We propose a weakly-supervised OOD detector (Figure 1) to identify skin images from different collections protocols and of unknown skin disease types, based on subset scanning [22] and ODIN [12]. We further evaluate algorithmic fairness of the proposed work across skin tones. We describe our proposed approach in detail next.

Subset scanning for OOD sample detection Given a pre-trained network C for skin disease classification, we apply subset scanning [22] on the activations in the intermediate layers of C to detect a subset S of OOD samples (see Algorithm 1). Subset scanning searches for the most anomalous subset $S^* = \arg\max_S F(S)$ in each layer, where the anomalousness is quantified by a scoring function $F(\cdot)$, such as a log-likelihood ratio statistic. When searching for this subset, an exhaustive search across all possible subsets is computationally infeasible as the number of subsets (2^N) increases exponentially with the number of nodes (N) in a layer. Instead, we utilize a scoring function that satisfies the Linear Time Subset Scanning [23] property, which guarantees that the highest-scoring subset of nodes in a layer are identified within $\approx N$ searches instead of 2^N searches. Following the literature on pattern detection [24, 22], we utilize non-parametric scan statistics (NPSS) [24] as our scoring function as it makes minimal assumptions on the underlying distribution of node activations.

We apply subset scanning on set of layers C_Y of C. For each layer $C_y \in C_Y$, we form a distribution of expected activations at each node using the known ID samples X_z, which were used during training and can also be referred as background data. Comparing this expected distribution to the node activations of each test or evaluation sample X_i, we can obtain p-values p_{ij} for each i^{th} test sample and j^{th} node of layer C_y. We can then quantify the anomalousness of the p-values by finding the subset of nodes that maximize divergence of the test sample activations from the expected. This yields $|C_Y|$ anomalous scores $S^*_{(C_y)}$ for each test sample. We expect OOD samples to yield higher anomalous scores S than ID samples, and detect OOD samples with simple thresholding. Note that the OOD detection is performed in an unsupervised fashion without any prior knowledge of the OOD samples.

ODIN and ODIN$_{low}$ Perturbations We also evaluated the impact of adding small perturbations, ODIN$_{low}$ and ODIN [12], to each test sample prior to subset scanning. ODIN involves two steps, input pre-processing and temperature scaling. In the first step, X_i is pre-processed by adding a small perturbation computed by back-propagating the gradient of the training loss with respect to X_i and weighted by parameter ϵ. This pre-processed X_i is then fed into the network, and temperature scaling with parameter τ is applied in the final softmax layer C_s. The two hyperparemters, ϵ and τ, are chosen so that the OOD detection performance of Softmax Score [13], the maximum value of the softmax layer output, is optimized. We modified ODIN and propose ODIN$_{low}$ with parameters τ_{low} and ϵ_{low} that leads to the lowest Softmax Score performance. As subset scanning is applied on the the inner layers of the network, using ODIN$_{low}$ helps improve OOD detection in the earlier layers of the network.

Algorithm 1: Proposed OOD detector

```
input : Background Image: $X_z \in D^{H_0}$, $M = |D^{H_0}|$,
        Evaluation Image: $X_i$, Training Dataset: $D_{train}$,
        Significance level: $\alpha_{\max}$.
output: $AUROC, F_1$ for $X_i$
C ← TrainSkinDiseaseClassifier ($D_{train}$);
$\hat{X}_z, \hat{X}_i$ ← AddODINlow ($X_z, X_i$);
for $C_y$ in $C$ do
    for $j \leftarrow 0$ to $|C_y|$ do
        $A^{H_0}_{zj}$ ← ExtractActivation ($C_y, \hat{X}_z$);
        $A_{ij}$ ← ExtractActivation ($C_y$,$\hat{X}_i$);
    $p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
    $p^s_{ij}$ ← SortAscending ($\{y < \alpha_{\max} \, \forall \, y \subseteq p_{ij}\}$);
    for $k \leftarrow 1$ to $|C_y|$ do
        $S_{(k)} = \{p_y \subseteq p^s_{ij} \forall y \in \{1, \ldots, k\}\}$;
        $\alpha_k = max(S_{(k)})$;
        $F(S_{(k)})$ ← NPSS ($\alpha_k$, k, k);
    $k^*_{(C_y)} \leftarrow \arg\max F(S_{(k)})$;
    $\alpha^*_{(C_y)} = \alpha_{k^*_{(C_y)}}$;
    $S^*_{(C_y)} = S_{(k^*_{(C_y)})}$;
return OODPerformance ($\sum_{C_y} S^*_{(C_y)}$)
```

Algorithmic Fairness of OOD detectors across skin tone We further evaluate algorithmic fairness of our OOD detector across skin tones, estimated by adopting an existing framework [25]. To this end, the non-diseased regions of a given skin image are segmented using Mask R-CNN [26], and individual typology angle is computed as $a = \arctan\left(\frac{L_\mu - 50}{b_\mu}\right) \times \frac{180^\circ}{\pi}$, where L_μ and b_μ are the average luminance and yellow values of non-diseased pixels. Using a, we stratify the samples into three Fitzpatrick skin tone categories, Light ($a > 41^\circ$), Intermediate ($28^\circ < a \leq 41^\circ$), and Dark ($a \leq 28^\circ$).

4. DATASETS

We validate the proposed framework using two datasets: ISIC 2019 [1, 27, 9] for samples of unknown diseases; and SD-198 [10] for samples from unknown collection protocols.

ISIC 2019 [1, 27, 9] dataset consists of $25,331$ dermoscopic images among eight diagnostic categories: *Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis, Benign keratosis, Dermatofibroma, Vascular lesion*, and *Squamous cell carcinoma*. As its test set is not available publicly, we set aside *Dermatofibroma* (DF) and *Vascular lesion* (VASC) samples during training, and utilize them as OOD samples of unknown diseases during testing.

SD-198 [10] dataset contain 198 diseases from different types of *eczema*, *acne* and *cancer*, totalling $6,584$ images. The images are collected via various devices, mostly digital cameras and mobile phones with higher levels of noise and varying illumination. We use this dataset for OOD samples from unknown collection protocols.

5. EXPERIMENTAL SETUP

We adopt DenseNet-121 [28] pre-trained on ImageNet [29] for skin disease classification and fine-tune it on ISIC 2019 [1] with seven

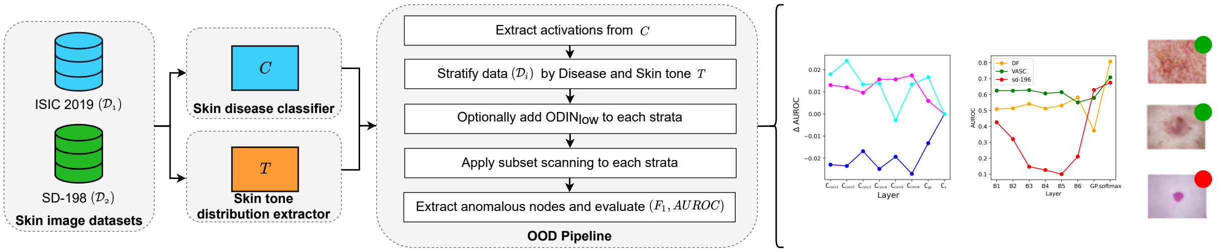

ISIC 2019 ($\mathcal{D}_1$)
SD-198 ($\mathcal{D}_2$)
Skin image datasets
C
Skin disease classifier
T
Skin tone distribution extractor
Extract activations from C
Stratify data ($\mathcal{D}_i$) by Disease and Skin tone T
Optionally add ODIN$_{low}$ to each strata
Apply subset scanning to each strata
Extract anomalous nodes and evaluate (F_1, $AUROC$)
OOD Pipeline
Δ AUROC
Layer
AUROC
DF
VASC
sd-196
Layer

Fig. 1. Block diagram of the proposed approach with a trained model for skin disease classification C over mentioned datasets ($\mathcal{D}_1$, $\mathcal{D}_2$).

Table 2. OOD detection performance for samples of unknown collection protocols, SD-198 [10]. Bold is best in each column.

Methods	AUROC	F_1
Softmax Score [13]	74.4 ± 1.7	71.0 ± 1.1
ODIN [12]	74.5 ± 1.6	70.8 ± 1.1
SS (C_s)	68.2 ± 1.4	71.3 ± 0.5
SS (C_{conv_1})	41.6 ± 1.8	68.1 ± 0.2
SS (C_{conv_1})+ODIN$_{low}$	85.4 ± 0.6	81.9 ± 0.6
SS (Sum All Layers)+ODIN$_{low}$	$\mathbf{91.0 \pm 0.8}$	$\mathbf{86.9 \pm 1.1}$

disease classes. Thus, we resize the last four fully connected layers to 512, 256, 128, and 7 nodes followed by a softmax. We use Adam [30] optimization on weighted cross-entropy loss with a learning rate of $1e^{-4}$ and a batch size of 40. We apply subset scanning across $|C_Y| = 8$ layers consisting of six convolutional layers ($C_{conv_1}, ..., C_{conv_6}$), global pooling layer (C_{gp}), and softmax layer (C_s). For ODIN$_{low}$, we use $\tau_{low} = 2$ and $\epsilon_{low} = 0.2$, which leads to the lowest Softmax Score performance (AUROC = 0.5) for both OOD use cases. To validate the detection of unknown disease samples, we use DF and VASC classes from ISIC-2019, consisting of 253 and 225 samples, respectively. For samples with different collection protocols, we extract ten sets of 260 samples from SD-198 and report their aggregate performance. We employ Area Under Receiver Operating Characteristic Curve (AUROC) and maximum F_1-score (F_1) as our OOD detection performance metrics.

6. RESULTS

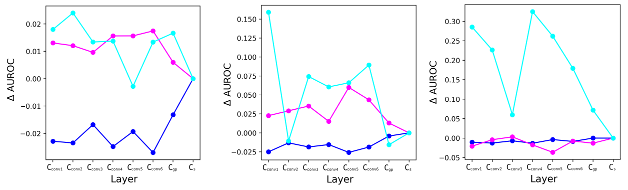
We first show the result of detecting OOD samples that are collected with different protocols (SD-198 [10]) in Table 2. In the top panel, we show the performance of the existing baselines, Softmax Score [13] and ODIN [12]. The remaining panels shows the result of our proposed approach - subset scanning (SS) with and without ODIN$_{low}$ noise. We achieve the best performance using the sum of subset scores across all layers $S^*_{(C_y)}$ with ODIN$_{low}$ (bolded).

Table 3 shows the performance of detecting OOD samples of unknown diseases (DF and VASC) that are unseen during training. Note that these OOD samples are from the same dataset as the training dataset. While Softmax Score [13] yields the best performance, subset scanning on the softmax layer C_s shows comparable performance. We see worse performances with ODIN$_{low}$ as these OOD samples are from the same dataset as ID samples and adding noise likely blurs the unique features present in each skin disease class.

Lastly, Figure 2 shows the change in AUROC of our proposed work with the stratification of OOD samples based on skin tone. While the samples of Light (blue) and Intermediate (magenta) skin tones show consistent performances throughout the eight layers C_Y that we consider, we see varying performances for those of Dark (cyan) skin tones. This instability of performance for Dark skinned samples may be partially because the network is trained on a datasets that heavily lacks samples of Dark skin tones. For instance, Dark skinned samples constitute only around 3.9% of DF and VASC samples and around 13% of SD-198 samples. This could also encourage OOD detectors to easily classify them to be out of distribution.

Table 3. OOD detection performance for samples of unknown disease types, DF and VASC [1]. Bold is best in each column.

Methods	AUROC		F_1	
	DF	VASC	DF	VASC
Softmax Score [13]	**80.9**	**73.2**	**76.5**	70.5
ODIN [12]	72.3	65.3	70.3	67.4
SS (C_s)	80.8	70.8	75.7	**72.3**
SS (C_{conv_1})	50.9	62.5	65.8	68.7
SS (C_{conv_1})+ODIN$_{low}$	47.6	39.8	65.9	67.1
SS (Sum All Layers)+ODIN$_{low}$	47.6	40.4	65.9	67.2



(a) SD-198 (b) DF (c) VASC

Fig. 2. Change in performance (AUROC) of our proposed OOD detector across C_Y layers with stratification into three skin-tone categories, Light (blue), Intermediate (magenta), and Dark (cyan).

7. CONCLUSION

We propose a weakly-supervised method to detect OOD skin images (collected in different protocols or from unknown disease types) using input perturbation and scanning of the activations in the intermediate layers of any pre-trained classifier. Our proposed method improves on the state-of-the-art for OOD samples that are collected from a different protocol, and it achieves competitive performance with the state-of-the-art in detecting samples of unknown diseases. We further stratify these OOD samples based on skin tone and observe imbalanced detection performance for Dark samples. Thus, future work aims to understand the reasons for such detection disparity across skin tones, e.g., lack of training representation or different manifestation of skin diseases.

8. COMPLIANCE WITH ETHICAL STANDARDS

This research uses human subject data made available in open access by the corresponding authors (ISIC 2019 [1, 27, 9] and SD-198 [10]) licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, and ethical approval was not required. No funding was received for this study.

9. REFERENCES

[1] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1710.05006, 2017.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] A. Gomolin, E. Netchiporouk, R. Gniadecki, and I. V. Litvinov, "Artificial intelligence applications in dermatology: Where do we stand?" *Front. Med.*, vol. 7, 2020.

[4] A. S. Adamson and A. Smith, "Machine learning and health care disparities in dermatology," *JAMA Derm.*, vol. 154, no. 11, pp. 1247–1248, 2018.

[5] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," arXiv:2001.08103, 2020.

[6] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Fairness of classifiers across skin tones in dermatology," in *Proc. Int. Conf. Med. Image Comp. Comp.-Assist. Interv.*, 2020, pp. 320–329.

[7] P. Zhang, Y. Zhong, and X. Li, "Melanet: A deep dense attention network for melanoma detection in dermoscopy images," 2019.

[8] S. A. A. Ahmed, B. Yanikoglu, E. Aptoula, and O. Goksu, "Skin lesion classification with deep learning ensembles in ISIC 2019," 2019.

[9] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "Bcn20000: Dermoscopic lesions in the wild," 2019.

[10] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Proc. Europ. Conf. Comput. Vis.*, 2016, pp. 206–222.

[11] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using loss balancing and ensembles of multi-resolution efficientnets," 2019.

[12] S. Liang, Y. Li, and R. Srikant, "Principled detection of out-of-distribution examples in neural networks," arXiv:1706.02690, 2017.

[13] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv:1610.02136, 2016.

[14] S. Bagchi, A. Banerjee, and D. R. Bathula, "Learning a meta-ensemble technique for skin lesion classification and novel class detection," in *CVPR Workshops*, June 2020.

[15] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, "Uncertainty estimation in deep neural networks for dermoscopic image classification," in *CVPR ISIC Skin Image Analysis Workshop*, 2020.

[16] A. G. C. Pacheco, A.-R. Ali, and T. Trappenberg, "Skin cancer detection based on deep learning and entropy to detect outlier samples," 2019.

[17] A. G. C. Pacheco, C. S. Sastry, T. Trappenberg, S. Oore, and R. A. Krohling, "On out-of-distribution detection algorithms with deep neural skin cancer classifiers," in *CVPR Workshops*, June 2020.

[18] M. Zaida, S. Ali, M. Ali, S. Hussein, A. Saadia, and W. Sultani, "Out of distribution detection for skin and malaria images," *CoRR*, vol. abs/2111.01505, 2021. [Online]. Available: https://arxiv.org/abs/2111.01505

[19] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[20] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 1050–1059.

[21] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with in-distribution examples and Gram matrices," 2019.

[22] C. Cintas, S. Speakman, V. Akinwande, W. Ogallo, K. Weldemariam, S. Sridharan, and E. McFowland, "Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error." in *IJCAI*, 2020, pp. 876–882.

[23] D. B. Neill, "Fast subset scan for spatial pattern detection," 2012.

[24] E. McFowland, S. Speakman, and D. B. Neill, "Fast generalized subset scan for anomalous pattern detection," *J. Mach. Learn. Res.*, vol. 14, no. 1, p. 1533–1561, Jan. 2013.

[25] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. F. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Estimating skin tone and effects on classification performance in dermatology datasets," 2019.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," arXiv:1803.10417, 2018.

[28] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," arXiv:1608.06993, 2016.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, 2009.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Repr.*, 2015.