Data-Driven Deep Supervision for Skin Lesion Classification

Suraj Mishra¹, Yizhe Zhang², Li Zhang³, Tianyu Zhang⁴, X. Sharon Hu¹, and Danny Z. Chen¹

¹ University of Notre Dame, Notre Dame, USA {smishra3,shu,dchen}@nd.edu
² Nanjing University of Science and Technology, Nanjing, China
³ Qingdao Women and Children’s Hospital of Qingdao University, Qingdao, China
⁴ University of Connecticut, Storrs, USA

Abstract. Automatic classification of pigmented, non-pigmented, and depigmented non-melanocytic skin lesions have garnered lots of attention in recent years. However, imaging variations in skin texture, lesion shape, depigmentation contrast, lighting condition, etc. hinder robust feature extraction, affecting classification accuracy. In this paper, we propose a new deep neural network that exploits input data for robust feature extraction. Specifically, we analyze the convolutional network’s behavior (field-of-view) to find the location of deep supervision for improved feature extraction. To achieve this, first we perform activation mapping to generate an object mask, highlighting the input regions most critical for classification output generation. Then the network layer whose layerwise effective receptive field matches the approximated object shape in the object mask is selected as our focus for deep supervision. Utilizing different types of convolutional feature extractors and classifiers on three melanoma detection datasets and two vitiligo detection datasets, we verify the effectiveness of our new method.

1 Introduction

Automatic classification of skin lesions such as pigmented/non-pigmented lesions concerning skin cancers [6,1,4,26,23,3] and depigmented and hypomelanic skin lesions concerning vitiligo (vitiligo vs vitiligo-type, e.g., pityriasis alba, nevus depigmentosus [29,13]) in RGB images has garnered lots of research attention in recent years. Convolutional neural network (CNN) based models have shown promising results in accurate classification of skin lesions. However, imaging variations in skin texture, lesion shape, depigmentation contrast, lighting condition, etc. hinder robust feature extraction, thus affecting classification accuracy.

For input data specific robust feature extraction, ensuring the discriminative-ness of the learned features in the hidden layers of a CNN is critical [15]. In [22], deep supervision was explored as an effective way to encourage feature maps at different layers to be directly predictive of the final output. By providing a companion objective using auxiliary classifiers [12], deep supervision generates feedback for a local output (at a hidden layer). The combined error (i.e., global
and local) is back-propagated, influencing the hidden layer update process and ensuring highly discriminative feature maps. But, where and how to apply deep supervision still remains an active research topic. Recently, Mishra et al. [17] studied data-driven deep supervision for medical image segmentation utilizing target object labels (or masks) as guidance for deep supervision. Yet, for image classification problems in general, the key issue of deep supervision location selection in a CNN becomes challenging due to the absence of object-level labels, since in classification problems, only image-level labels are normally available.

Previous work [15,17] exploited the observation that if the receptive field (RF) or field-of-view of a segmentation CNN matches in size the target object size in the input images, then the segmentation accuracy can be improved. Relative to the CNN’s RF, smaller objects are lost in the network’s sub-sampling operations while robust global features of larger objects are not well captured due to a smaller RF [17]. We extend the observation/technique of deep supervision to tackle various skin lesion classification tasks, proposing data-driven deep supervision. Specifically, for deep supervision location selection, we utilize the input domain information (target object sizes) to identify the layer with preeminent contribution to feature extraction, and select this layer as the target location for deep supervision as auxiliary classification. Hence, as in [17], we employ an RF based approach to determine the preeminent layer in a classification CNN. But, note that for image classification problems, object-level labels are usually not available (only image-level labels, e.g., {yes, no}, are given). Thus, a difficulty to our approach is how to devise an effective method to estimate the sizes of objects in input images that possibly influence the classification results.

Each convolutional (conv) layer of a CNN extracts local features from a region of an input image to generate its output. The RF of the CNN is defined as the input region that directly affects the CNN’s output [11]. By analyzing the near-Gaussian distribution of RF, Luo et al. [14] identified the unresponsive boundary regions of RF as non-output-affecting. The actual output-affecting region of the input image, which is smaller than the theoretical RF, is defined as the effective receptive field (ERF) of the network [14]. Intuitively, convolutional filters extract improved features when the target object size (called morphological size) matches the filters’ ERF and the target lies at the center of the ERF [17].

From the viewpoint of ERF matching, for skin lesion classification using a deep CNN, we need to address two challenges: (1) determining the characteristic size of objects in the input images that the CNN is applied to; (2) determining the CNN’s ERF, whose size should match the characteristic object size. For the first challenge, we explore activation mapping to determine the morphological size of the input objects (approximate regions of interest in input images which have significant contribution to output generation). Since such regions of interest can vary largely in size, for the second challenge, instead of using the fixed (or default) network ERF of a deep CNN, we utilize the layer-wise ERF (LERF) [15]. LERF represents the ERF of each conv layer in the CNN with respect to the image input. The conv layer with an LERF similar to the dataset-specific approximate object size is used as the target layer for deep supervision.
Fig. 1. Our proposed framework for data-driven deep supervision for skin lesion classification. The major components of our framework are (1) LERF determination, (2) morphological object size approximation, and (3) deep supervision employment, shown in blue, orange, and red background, respectively. Specifically, utilizing activation maps, a morphological object mask is generated. The network layer whose LERF size well matches the morphological object size is chosen for deep supervision.

In data-driven deep supervision for skin lesion classification, we first approximate the morphological input object size from image-level labels using activation mapping. Then we introduce deep supervision from the layer of the network whose LERF size matches (as well as possible) the morphological object size. We validate our approach on three pigmented/non-pigmented lesion classification datasets concerning skin cancer (ISIC 2016 [6], ISIC 2017 [1], and ISIC 2018 [2]) and two depigmented/hypopigmented lesion classification datasets concerning vitiligo (vitiligo vs vitiligo-type, e.g., pityriasis alba, nevus depigmentous). For the public ISIC 2016, ISIC 2017, and ISIC 2018 datasets, we achieve comparable scores with the challenge winners. For both the vitiligo datasets [29], we outperform existing methods.

2 Method

The overall framework of our proposed data-driven deep supervision for skin lesion classification is shown in Fig. 1. The major components of our framework are (1) LERF determination, (2) morphological object size approximation, and (3) deep-supervision employment. Section 2.1 and Section 2.2 highlight LERF determination and morphological object size approximation, respectively. We then show how to use the information thus obtained for data-driven deep supervision in Section 2.3, along with some details of deep supervision implementation.

2.1 Layer-wise Effective Receptive Field (LERF)

We take a partial derivative based approach to measure the influence on an input image region by an output node (a pixel) of a conv layer as in [14,17]. For a classification CNN with \( N \) conv layers, \( LERF_h \) is associated to layer \( h \)
Fig. 2. (a) Gaussian distribution of receptive field generated using the VGG13 architecture [21] with linear activations. (b) The boundary of RF is shown in red square. The boundary of ERF is shown in dashed white lines. Plots are generated for the last layer of the feature extractor (L34). (c)-(e) LERF plots are for L28, L24, and L21. (f) The extracted convolutional hierarchical features are average-pooled and classified.

(h = 1, 2, ..., N). Our objective is to determine $\frac{\partial q_{i,t,j_t}}{\partial p_{i,j}}$, where $p_{i,j}$ is an input node (pixel) and $q_{i_t,j_t}$ is the target node at the output of the conv layer $h$ for which the influence measurement is being performed. Following [17], the region of influence (i.e., $\frac{\partial q_{i_t,j_t}}{\partial p_{i,j}}$) is determined using backpropagation (as all non-zero gradients on the input image). In Fig. 2, LERFs associated to different layers of a representative CNN network are shown. Any node $q_{i_t,j_t}$ of a target layer can be selected for backpropagation since different nodes only generate translational shifts in the output (see Fig. 2).

To extract the LERF size from $\frac{\partial q_{i_t,j_t}}{\partial p_{i,j}}$, first we threshold (1 if $\frac{\partial q_{i_t,j_t}}{\partial p_{i,j}} > th$; else 0) to remove the unresponsive boundary regions. Then, the pixel count of the active region is square-rooted to approximate the size of the LERF. Such a size approximation is inspired by the near-Gaussian distribution of ERF [17]. Following [14], a $th = 0.0455$ (i.e., a range of two standard deviations of the Gaussian distribution, 4.55% of the maximum value, i.e., 1 at the center) is used in LERF calculation. In Fig. 2, the convex hull of the LERF pixels (the smallest convex area covering all the pixels) is marked with a closed white dotted curve. With non-linear activations (i.e., ReLU), this convex hull has a shape of a deformed square or an evolved ellipse. In the absence of network non-linearity, LERF has a circular shape (see Fig. 2). Hence, a square root based operation is used to approximate such a deformed shape of LERF with network non-linearity.

2.2 Object Size Approximation Using Activation Mapping

We explore an activation mapping based morphological object size approximation. Consider a classification CNN (generating an output label for an input image), which contains a stack of conv, activation, and sub-sampling layers for feature extraction. CNN-extracted features are average-pooled and fed to a fully connected (FC) classifier for final output generation (see Fig. 2(f)).

For an input image, propagating through the CNN, let $f_k(i, j)$ represent the activation of a unit/channel/kernel $k$ in the last conv layer at spatial location $(i, j)$. Then performing average pooling for unit $k$ results in $F_k = \sum_{i,j} f_k(i, j)$. Ignoring the bias term, the FC classifier layer with weights $w^c_k$ associated to a class $c$ generates output $\sum_k w^c_k F_k$ for class $c$. This output is used for loss calculation after a softmax-type operation. By plugging the value of $F_k$ into the
Fig. 3. (a) A CNN with extracted conv features being pooled and classified for final output generation. (b) An activation map generated for a predicted class yes. The generated activation map is thresholded with $\delta = 0.5$ to generate a summary-mask. (c)-(d)-(e)-(f) Example images, activation maps in gray, activation maps in color, and summary-masks, respectively. (g) Bounding boxes for the central regions. (h)-(i) Object masks shown on activation maps and input images, respectively.

class score, we obtain $\sum_k w^c_k \sum_{i,j} f_k(i,j) = \sum_k \sum_{i,j} w^c_k f_k(i,j)$. In [31], the term $\sum_k w^c_k f_k(i,j)$ is defined as the activation map ($A_c$) for class $c$ (see Fig. 3(b)), indicating the importance of the activation leading to the classification of an image for class $c$. As activation mapping indicates the importance of activations on an input image, we utilize it for object mask generation.

For object mask generation, we utilize the activation map $A_c$ and reshape it to the input image size using cubic-interpolation to obtain $A'_c$. The reshaped map $A'_c$ is then thresholded with a threshold $\delta'$ (i.e., $A'_c(i,j) = 0$ if $A'_c(i,j) < \delta'$; else $A'_c(i,j) = 1$). We use $\delta' = 0.9$ to determine the central region. The smallest bounding box containing the central region is computed (see Fig. 3(g)). Then we perform a region-growing type process to increase the size of the bounding box (equally in four directions) till we reach a pixel with a 50% probability value ($\delta = 0.5$) in the activation map. A box in Fig. 3(g) is increased in size till it touches the boundary of the summary mask (see Fig. 3(f)), and the result is shown in Fig. 3(h). Such processes are based on the special and stable properties of skin images (similar appearance, captured in similar conditions, using specific devices, etc). A common target lesion in skin images is a single connected region lying around the center of an image. With our approach, the most critical regions associated with the classification of the images in Fig. 3(c) are shown in Fig. 3(i).

In Fig. 3(h)-(i), example object masks are shown for some vitiligo images. It is interesting to note that the transition regions have the highest impact on final output generation. As in Sec. 2.1, square root of the pixel count of an object mask (red bounding box regions in Fig. 3(i)) is used as an approximate size of the object mask. The average size of all such object masks over sampled training set images is used as the morphological object size ($\text{Obj}$) of a specific dataset.

### 2.3 Deep Supervision Employment

To apply deep supervision, we use $\text{LERF}_h$'s (computed in Section 2.1) and $\text{Obj}$ (computed in Section 2.2) to find a target layer $L_{\text{target}}$, i.e., the conv layer with
preeminent contribution to feature extraction. $L_{\text{target}}$ is taken as the layer $h^*$ with $|\text{Obj} - LERF_{h^*}| \leq |\text{Obj} - LERF_h|, \forall h \in \{1, \ldots, N\}$. Intuitively, $L_{\text{target}}$ is the conv layer in an $N$-layer CNN whose $LERF$ size best matches the $\text{Obj}$ size. To perform deep supervision using the output of $L_{\text{target}}$, we need to classify the features associated with $L_{\text{target}}$. The classification scores are then used for the auxiliary loss calculation. Our total loss is computed as $Loss_{\text{total}} = Loss_{\text{main}} + Loss_{\text{aux}}$, where $Loss_{\text{main}}$ is the main loss and $Loss_{\text{aux}}$ is the auxiliary loss.

Remarks on Implementation. Here, we discuss two aspects related to deep supervision implementation: (1) auxiliary classifier design; (2) transfer learning.

For deep supervision employment, conv features are processed by a classifier. The classified output is then used to compute the auxiliary loss. In this work, we explore two types of classifier architectures for auxiliary classifier design. (i) In a VGG-type classifier [21], multiple FC layers are used along with activation and dropout layers. In contrast, (ii) a ResNet-type classifier [7] is light weight and uses a single FC layer as the classifier (details in Supplemental Material).

Transfer learning plays a critical role (better/quicker optimization) in image classification by initiating the training process with a previously trained model [20], and has shown to be effective with improved classification accuracy. In data-driven deep supervision, an auxiliary loss is introduced from the preeminent layer of the network. Such model modifications encourage us to adopt a super-model based encapsulation approach, in which the base model is initialized by transfer learning and the auxiliary branch is initialized randomly (He-initialization [8]). Benefits of transfer learning are shown in Section 3.

3 Experiments and Results

Datasets. Our experiments use the following datasets. ISIC 2016 [6]: It contains 900 training and 379 test images. We randomly divide the original training set into 649 images for training and 251 images for validation. ISIC 2017 [1]: It contains 2000 training, 150 validation, and 600 test images. Experiments are performed for task-3A, i.e., melanoma detection. ISIC 2018 [2]: It contains 10,015 training, 193 validation, and 1512 test images with external evaluation. Considering the smaller size of the validation set, we randomly select 999 images ($\approx 10\%$) from the training set and add them to the validation set (9016 train and 1192 validation images). Vitiligo (in-house): It contains 2188 images (1227 training, 308 validation, and 653 test images). The in-house dataset consists of images from retrospective consecutive outpatients obtained by the dermatology department of Qingdao Women and Children’s Hospital (QWCH) in China. For each patient with suspected vitiligo (e.g., pityriasis alba, hypopigmented nevus), clinical photographs of the affected skin areas were taken by medical assistants using a point-and-shoot camera (as described in [29]). Vitiligo (public) [29]: It contains 1341 images (672 training, 268 validation, and 401 test images).

Experimental Setup. We resize all the images of each dataset to 224×224. To reduce overfitting on a small training set, data augmentation is performed using random flipping and cropping. In cropping, a random portion of an image
Table 1. Lesion classification results on skin image datasets. Proposed*/Proposed — without/with deep supervision.

| Method               | ISIC 2016 Dataset | ISIC 2017 Dataset | Vitiligo (in-house) Dataset | Vitiligo (public) Dataset |
|----------------------|-------------------|-------------------|-----------------------------|--------------------------|
|                      | AUC | Acc | Sen | Spe | AUC | Acc | Sen | Spe | AUC | Acc | Sen | Spe | AUC | Acc | Sen | Spe |
| VGG-16 [21]          | 0.826 | 0.826 | 0.413 | 0.928 |      |      |      |      |      |      |      |      |      |      |      |      |
| GoogleNet [22]       | 0.801 | 0.847 | 0.507 | 0.931 |      |      |      |      |      |      |      |      |      |      |      |      |
| DRN-50 [27]          | 0.783 | 0.855 | 0.547 | 0.931 |      |      |      |      |      |      |      |      |      |      |      |      |
| Gutman et al. [6]    | 0.804 | 0.855 | 0.507 | 0.941 |      |      |      |      |      |      |      |      |      |      |      |      |
| ARDT-DenseNet [25]   | 0.837 | 0.857 | 0.816 | 0.931 |      |      |      |      |      |      |      |      |      |      |      |      |
| SDL [28]             | 0.829 | 0.857 |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed* - VGG13    | 0.850 | 0.852 | 0.334 | 0.961 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed - VGG13     | 0.857 |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed* - RES18    | 0.817 | 0.850 | 0.400 | 0.961 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed - RES18     | 0.850 | 0.842 | 0.440 | 0.941 |      |      |      |      |      |      |      |      |      |      |      |      |
| VGG13 [21]           | 0.851 | -    | 0.791 | 0.913 |      |      |      |      |      |      |      |      |      |      |      |      |
| ResNet [7]           | 0.840 | -    | 0.775 | 0.902 |      |      |      |      |      |      |      |      |      |      |      |      |
| Dermatologists [29]  | -    | -    | 0.611 | 0.999 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed* - VGG13    | 0.852 | 0.882 | 0.797 | 0.977 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed - VGG13     | 0.851 | 0.904 | 0.777 | 0.941 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed* - RES18    | 0.897 | 0.865 | 0.662 | 0.932 |      |      |      |      |      |      |      |      |      |      |      |      |
| Proposed - RES18     | 0.920 | 0.893 | 0.757 | 0.933 |      |      |      |      |      |      |      |      |      |      |      |      |

Results and Discussions. Table 1 shows the results. For each dataset, the Proposed* or Proposed row gives our results obtained without or with deep supervision. The backbone CNN networks used in each experiments (i.e., VGG13 and RES18) are shown. Applying data-driven deep supervision indeed improves classification accuracy. On the ISIC 2016 dataset, our method exceeds the results...
Fig. 4. (a)-(e)-(i) Example images. (b)-(f)-(j) Activation maps generated without deep supervision. (c)-(g)-(k) Activation maps by deep-supervision (main branch). (d)-(h)-(l) Activation maps by deep-supervision (aux branch).

Table 2. Ablation study analysis. C: Classifier type (R(ResNet-type)/V(VGG-type)); DS: Deep supervision location; TL: Transfer learning (with(✓)/without(✗)).

| Method     | Backbone | ISIC 2016 Dataset | Vitiligo (in-house) Dataset | Vitiligo (public) Dataset |
|------------|----------|-------------------|----------------------------|---------------------------|
|            |          | AUC   | Acc   | Sen  | Spe  | AUC   | Acc   | Sen  | Spe  | AUC   | Acc   | Sen  | Spe  | AUC   | Acc   | Sen  | Spe  | AUC   | Acc   | Sen  | Spe  |
| VGG13      | R        | ✓     | 0.830 | 0.852 | 0.334 | 0.980 | 0.852 | 0.882 | 0.797 | 0.907 | 0.970 | 0.973 | 0.992 | 0.951 |          |        |        |        |        |        |        |        |
| VGG13      | R L28    | ✓     | 0.837 | 0.877 | 0.387 | 0.990 | 0.931 | 0.904 | 0.777 | 0.941 | 0.998 | 0.988 | 0.996 | 0.975 |          |        |        |        |        |        |        |        |
| VGG13      | R L24    | ✓     | 0.830 | 0.839 | 0.293 | 0.974 | 0.924 | 0.907 | 0.701 | 0.941 | -     | -     | -     | -     |          |        |        |        |        |        |        |        |
| VGG13      | R L24 & L28 | ✓     | 0.834 | 0.847 | 0.320 | 0.977 | 0.931 | 0.897 | 0.716 | 0.950 | -     | -     | -     | -     |          |        |        |        |        |        |        |        |
| VGG13      | R L31    | ✓     | 0.838 | 0.852 | 0.387 | 0.967 | 0.936 | 0.891 | 0.811 | 0.915 | 0.907 | 0.973 | 0.983 | 0.957 |          |        |        |        |        |        |        |        |
| VGG13      | V L28    | ✓     | 0.715 | 0.810 | 0.227 | 0.954 | 0.809 | 0.860 | 0.703 | 0.915 | 0.892 | 0.905 | 0.962 | 0.821 |          |        |        |        |        |        |        |        |
| VGG13      | V L28    | ✓     | 0.744 | 0.829 | 0.213 | 0.980 | 0.842 | 0.862 | 0.804 | 0.879 | 0.921 | 0.928 | 0.954 | 0.889 |          |        |        |        |        |        |        |        |
| VGG13      | V L28    | ✓     | 0.840 | 0.847 | 0.400 | 0.957 | 0.917 | 0.891 | 0.703 | 0.947 | 0.998 | 0.980 | 0.992 | 0.963 |          |        |        |        |        |        |        |        |

of the challenge winner [6]. On ISIC 2017, our results are comparable to the best performing models in the leader-board with published results [4,26,23,3]. On the ISIC 2018 dataset, external evaluation is performed and we receive the results as an “overall score” (aggregated over all the 7 classes), i.e., no Acc, Sen, Spe, and AUC results for it. Our Proposed*-VGG13 achieves an overall score of 0.664 (comparable to the challenge rank 50 [18]); our Proposed-VGG13 yields an overall score of 0.701 (comparable to the challenge rank 10 [5] among the published results without using external data, with the best reported score being 0.845 [32]). For our Proposed*-RES18, an overall score of 0.626 is received. With supervision, our Proposed-RES18 achieves an overall score of 0.661. These highlight the benefits of our deep supervision. To further examine the impact of data-driven deep supervision on skin lesion classification, we compare changes in activation maps with deep supervision. In Fig. 4, changes in activations due to deep supervision for example images are shown.

Ablation Study. Using ISIC 2016, Vitiligo (in-house), and Vitiligo (public) as example cases, we conduct an ablation study to examine the contributions of deep supervision location selection, classifier architecture, and transfer learning. Table 2 shows the results. Deep Supervision Location Selection: Contribution of our deep supervision is shown by the DS column in Table 2. In addition to data-driven deep supervision (at L28 in VGG13), experiments are conducted by applying deep supervision at different layers (L24, L24 & L28, and L31). On the Vitiligo (in-house) dataset, better AUC and specificity are observed for deep supervision with the L31, and L24 & L28 case, respectively. Transfer Learn-
ing: Contribution of transfer learning is shown in the TL column of Table 2. In the absence of transfer learning, there is a significant drop in all accuracy metrics. **Classifier Architecture:** Experiments using VGG-type classifier architectures are discussed in Section 2.3. Both the classifier architectures generate comparative results (in column C of Table 2).

4 Conclusions

In this paper, we proposed a new data-driven deep supervision approach for skin lesion classification. Utilizing CNNs’ layer-wise ERF information and input object size approximated by activation mapping, deep supervision location is selected. Experiments on various datasets verify the effectiveness of our approach.

In future work, we plan to improve the robustness of our proposed method by substituting CAM [31] with improved frameworks such as grad-CAM [19] and score-CAM [24]. Further, the single target lesion assumption is a limitation of the proposed method. We plan to extend our method to multiple target objects of different sizes based on the data-driven deep supervision segmentation method for multiple object sizes as in [17,30,16].

Acknowledgement.
This work was supported in part by NSF grant CCF-1617735.

References

1. Codella, N.C.F., et al.: 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). In: ISBI 2018. pp. 168–172 (2018). https://doi.org/10.1109/ISBI.2018.8363547
2. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). CoRR abs/1902.03368 (2019), http://arxiv.org/abs/1902.03368
3. Galdran, A., et al.: Data-driven color augmentation techniques for deep skin image analysis. CoRR abs/1703.03702 (2017), http://arxiv.org/abs/1703.03702
4. González-Díaz, I.: Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. CoRR abs/1703.01976 (2017), http://arxiv.org/abs/1703.01976
5. Goyal, M., Rajapakse, J.C.: Deep neural network ensemble by data augmentation and bagging for skin lesion classification. CoRR abs/1807.05496 (2018), http://arxiv.org/abs/1807.05496
6. Gutman, D.A., et al.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). CoRR abs/1605.01397 (2016), http://arxiv.org/abs/1605.01397
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015), http://arxiv.org/abs/1512.03385
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR abs/1502.01852 (2015), http://arxiv.org/abs/1502.01852
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243

10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2015), http://arxiv.org/abs/1412.6980

11. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1(4), 541–551 (1989)

12. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-Supervised Nets. In: International Conf. on Artificial Intelligence and Statistics. pp. 562–570 (2015)

13. Liu, J., Yan, J., Chen, J., Sun, G., Luo, W.: Classification of vitiligo based on convolutional neural network. In: Sun, X., Pan, Z., Bertino, E. (eds.) Artificial Intelligence and Security. pp. 214–223 (2019)

14. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: NeurIPS. p. 4905–4913 (2016)

15. Mishra, S., Chen, D.Z., Hu, X.S.: A data-aware deep supervised method for retinal vessel segmentation. In: ISBI. pp. 1254–1257 (2020)

16. Mishra, S., Chen, D.Z., Hu, X.S.: Objective-dependent uncertainty driven retinal vessel segmentation. In: ISBI. pp. 453–457 (2021). https://doi.org/10.1109/ISBI48211.2021.9433774

17. Mishra, S., Zhang, Y., Chen, D.Z., Sharon Hu, X.: Data-driven deep supervision for medical image segmentation. IEEE Transactions on Medical Imaging (2022). https://doi.org/10.1109/TMI.2022.3143371

18. Pal, A., Ray, S., Garain, U.: Skin disease identification from dermoscopy images using deep convolutional neural network. CoRR abs/1807.09163 (2018), http://arxiv.org/abs/1807.09163

19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)

20. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging 35(5), 1285–1298 (2016)

21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)

22. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)

23. Vasconcelos, C.N., Vasconcelos, B.N.: Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. CoRR abs/1702.07025 (2017), http://arxiv.org/abs/1702.07025

24. Wang, H., Du, M., Yang, F., Zhang, Z.: Score-CAM: Improved visual explanations via score-weighted class activation mapping. CoRR abs/1910.01279 (2019), http://arxiv.org/abs/1910.01279

25. Wu, J., Hu, W., Wen, Y., Tu, W., Liu, X.: Skin lesion classification using densely connected convolutional networks with attention residual learning. Sensors 20(24) (2020). https://doi.org/10.3390/s20247080, https://www.mdpi.com/1424-8220/20/24/7080

26. Yang, X., Zeng, Z., Yeo, S.Y., Tan, C., Tey, H.L., Su, Y.: A novel multi-task deep learning model for skin lesion segmentation and classification. CoRR abs/1703.01025 (2017), http://arxiv.org/abs/1703.01025
27. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Transactions on Medical Imaging 36(4), 994–1004 (2017)
28. Zhang, J., Xie, Y., Wu, Q., Xia, Y.: Medical image classification using synergic deep learning. Medical Image Analysis 54, 10–19 (2019)
29. Zhang, L., Mishra, S., et al.: Design and assessment of convolutional neural network based methods for vitiligo diagnosis. Frontiers in Medicine 8 (2021). https://doi.org/10.3389/fmed.2021.754202
30. Zhang, Y., Ying, M., Chen, D.: Decompose-and-integrate learning for multi-class segmentation in medical images. In: MICCAI. pp. 641–650 (2019)
31. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. CVPR (2016)
32. Zhuang, J., Li, W., Manivannan, S., Wang, R., Zhang, J., Pan, J., Jiang, G., Yin, Z.: Skin lesion analysis towards melanoma detection using deep neural network ensemble. ISIC Challenge 2018(2), 1–6 (2018)
Supplemental Material for Submission #1816
“Data-Driven Deep Supervision for Skin Lesion Classification”

No Author Given
No Institute Given

Fig. 1. (a) A VGG-type classifier with multiple fully-connected layers. (b) A ResNet-type classifier with a single fully-connected layer. (c) Our super model based encapsulation approach for transfer learning implementation for the modified network architecture with deep supervision.

```python
class VGG16DS(nn.Module):
    def __init__(self):
        super().__init__()
        #Pretrained PyTorch Model <--- Transfer Learning
        self.model = models.vgg16_bn(pretrained=True)
        #Customizing the original model for a specific task
        num_ftrs = self.model.classifier[6].in_features
        self.model.classifier[6] = nn.Linear(num_ftrs, num_classes)
        #Auxiliary Classifier
        self.avpoo1 = nn.AdaptiveAvgPool2d((1,1))
        self.classifier1 = nn.Linear(512, num_classes)

    def forward(self, x):
        #Main output is generated
        main = self.model(x)
        #Selecting the location of Deep Supervision
        second = self.model.features[-1](x)
        #Auxiliary output is generated
        avpoo2 = self.avpoo2(second)
        avpoo2_flat = torch.flatten(avpoo2, 1)
        aux = self.classifier2(avpoo2_flat)
        return (main, aux)
```

Fig. 2. PyTorch implementation of our encapsulated model with auxiliary branch.
Table 1. Approximated object size (Obj) for different datasets.

|                | ISIC 2016 | ISIC 2017 | ISIC 2018 | Vitiligo (in-house) | Vitiligo (public) |
|----------------|-----------|-----------|-----------|---------------------|-------------------|
| Obj            | 67.19     | 71.53     | 66.07     | 67.86               | 66.31             |

Fig. 3. VGG13 layer arrangement.