PSENet: Psoriasis Severity Evaluation Network

Yi Li,[1*] Zhe Wu,[1,3*] Shuang Zhao,[2*] Xian Wu,[3] Yehong Kuang,[2]
Yangtian Yan,[3] Shen Ge,[3] Kai Wang,[3] Wei Fan,[3] Xiang Chen,[2†] Yong Wang[1†]
[1]School of Automation, Central South University
[2]Department of Dermatology, Xiangya Hospital, Central South University
[3]Tencent Medical AI Lab
{liyi1002, wuzhe950818, shuangxy, ywang}@csu.edu.cn, {yh_927, chenxiangck}@126.com
{kevinxwu, yangtianyan, shenge, ironswang, davidwfan}@tencent.com

Abstract

Psoriasis is a chronic skin disease which affects hundreds of millions of people around the world. This disease cannot be fully cured and requires lifelong caring. If the deterioration of Psoriasis is not detected and properly treated in time, it could cause serious complications or even lead to a life threat. Therefore, a quantitative measurement that can track the Psoriasis severity is necessary. Currently, PASI (Psoriasis Area and Severity Index) is the most frequently used measurement in clinical practices. However, PASI has the following disadvantages: (1) Time consuming: calculating PASI usually takes more than 30 minutes which poses a heavy burden on dermatologists; and (2) Inconsistency: due to the complexity of PASI calculation, different or even the same dermatologist could give different scores for the same case. To overcome these drawbacks, we propose PSENet which applies deep neural networks to estimate Psoriasis severity based on skin lesion images. Different from typical deep learning frameworks for image processing, PSENet has the following characteristics: (1) PSENet introduces a score refine module which is able to capture the visual features of skin at both coarse and fine-grained granularities; (2) PSENet uses siamese structure in training and accepts pairwise inputs, which reduces the dependency on large amount of training data; and (3) PSENet can not only estimate the severity, but also locate the skin lesion regions from the input image. To train and evaluate PSENet, we work with professional dermatologists from a top hospital and spend years in building a golden dataset. The experimental results show that PSENet can achieve the mean absolute error of 2.21 and the accuracy of 77.87% in pair comparison, outperforming baseline methods. Overall, PSENet not only relieves dermatologists from the dull PASI calculation but also enables patients to track Psoriasis severity in a much more convenient manner.

Introduction

Psoriasis is a chronic inflammatory skin disease which affects about 2%-3% of the population worldwide[1] . It is an immune system disease and cannot be fully cured. As a result, many patients have to suffer from Psoriasis throughout their entire lives. But if treated properly and timely, patients can still maintain a relatively high life quality. For Psoriasis with different severities, different therapies can be chosen accordingly. For example, balm, radiation, and biologic are suitable treatments for mild, medium, and serious conditions, respectively. However if the deterioration of Psoriasis is not detected and handled in time, it could cause serious complications, such as diabetes and heart failure (Berth-Jones et al. 2006; Zhou et al. 2015). Therefore, tracking the progress of Psoriasis is of great importance which requires a quantitative indicator.

In current clinical practice, Psoriasis Area and Severity Index (PASI) (Berth-Jones et al. 2006) is the most frequently used indicator. PASI mainly evaluates the severity of skin lesions from four aspects: the redness of erythema, the thickness of induration, the scaling of desquamation, and the lesion area ratio, resulting in a score ranging from 0 to 72. One drawback of PASI is that it is manually calculated by dermatologists and usually takes more than 30 minutes per patient (Berth-Jones et al. 2006). However, there is a serious shortage of dermatologists in less developed and developing countries. For example, in Africa, there are few dermatologists in many countries; and in China, the ratio between dermatologists and patients is 1:70000 (Zhou et al. 2015). Even for developed countries, considering the large number of Psoriasis patients, calculating PASI is still a heavy burden for dermatologists. Another drawback of PASI is the inconsistency. Due to the complexity of PASI calculation, different dermatologists could come up with different scores for the same case. Even the same dermatologist may give different scores for the same case. Such inconsistency could mislead the judgement of Psoriasis severity.

To overcome the above disadvantages of PASI, we propose PSENet which applies deep neural network to estimate the severity. PSENet takes clinical images of Psoriasis as input and generates a numeric score to measure the severity. It has the following two components: (1) Score refine module: this module simulates the process how PASI evaluates the severity. On the one hand, this module locates skin lesions and estimates their sizes. On the other hand, this module evaluates the abnormality degree of skin lesions. By combining these two results, this module can output a numeric score to estimate the severity with high robustness. This module is deployed at different depths of the deep neural network. In

*These authors contributed equally. This work has been done when Zhe Wu was with Tencent Medical AI Lab as intern.
†Corresponding authors

[1]http://www.worldpsoriasisday.com/

this manner, it is able to capture the visual features of skin from coarse to fine-grained granularities; and (2) siamese structure: PSENet adopts siamese structure (Chopra et al. 2005) in the training stage to model severity evaluation as a metric learning task. In addition to estimate the score of a single image, PSENet can take two different images as an input sample and predict the difference between their severities. With such a training strategy, for a dataset with N images, we can acquire $N*(N-1)/2$ pairs of images as training samples.

Since PSENet is a supervised model, its performance highly depends on the quality and quantity of the training data. To this end, we collaborate with a world famous dermatologist team and spend years in building a golden dataset. Each instance is verified pathologically and its severity score is given by experienced dermatologists.

To evaluate PSENet, we introduce two metrics: mean absolute error (MAE) and pair accuracy (PA): MAE focuses on the absolute score difference between PSENet and the dermatologists, while PA focuses on the relative ranking order between two images given by PSENet and dermatologists. As shown in the experimental results, PSENet achieves the MAE of 2.21 (in the range of 72) and the PA of 77.87%, outperforming all baseline models.

Compared with manually calculated PASI, PSENet can produce consistent measurement in a much more time-effective manner. It can be easily deployed as a web service or an APP on smart phones. For patients, they only need to upload a skin lesion picture and can obtain a severity score in seconds; and for dermatologists, PSENet can relieve them from the dull PASI scoring. In this way, patients can track the progress of Psoriasis without going to hospital. This is especially beneficial for patients in places where dermatologists are hard to reach.

PASI Description

Due to the importance of Psoriasis severity estimation, two metrics, i.e., PASI and Psoriasis Global Assessment (PGA), have been proposed (Langley and Ellis 2004), in which PASI is the most frequently used one in clinical treatment. For a patient, the calculation process of PASI is as follows:

(1) Dermatologists collect clinical images from four different body parts of a patient, including head, torso, upper limbs, and lower limbs.

(2) For each body part, dermatologists use three indicators to evaluate the seveirty of Psoriasis: the redness of erythema, the thickness of induration, and the scaling of desquamation. Each indicator is represented by a score ranging from 0 to 4. The higher the score, the more severe the condition. Then these three scores are added up to represent the general severity of this body part.

(3) Dermatologists manually estimate the proportion of skin lesion area to normal skin area, resulting in a corresponding score (range from 0 to 6 for 0% to 100%).

(4) After the four scores for each body part in Step 2 and Step 3 are estimated, dermatologists use equations (1) and (2) to calculate the final score of PASI.

$$PASI_{part_i} = (S_{ery_{part_i}} + S_{ind_{part_i}} + S_{des_{part_i}}) \times A_{part_i} \quad (1)$$

$$PASI_{patient} = \sum_{i}^{parts} W_{part_i} \times PASI_{part_i} \quad (2)$$

where $part_i \in \{head, torso, upper\ limbs, lower\ limbs\}$ denotes the ith body part; $W_{part_i} \in \{0.1, 0.3, 0.2, 0.4\}$ denotes the corresponding weight; $S_{ery_{part_i}}$, $S_{ind_{part_i}}$, and $S_{des_{part_i}}$ denote the severity scores for the redness of erythema, the thickness of induration, and the scaling of desquamation, respectively, and A_{part_i} denotes the proportion score. Clinically, $S_{ery_{part_i}}$, $S_{ind_{part_i}}$, and $S_{des_{part_i}}$ range from 0 to 4; A_{part_i} ranges from 0 to 6; and $PASI_{patient}$ ranges from 0 to 72 (George, Aldeen, and Garnavi 2017).

Such a scoring process has two drawbacks. First, dermatologists have to estimate 4 scores $S_{ery_{part_i}}$, $S_{ind_{part_i}}$, $S_{des_{part_i}}$ and A_{part_i} for each body part; therefore, there are $4 * 4$ variables in total for a single patient. The estimation heavily depends on the expertise of dermatologists, and different dermatologists are likely to give different estimations. Even for the same dermatologist, the estimation for the same case may be varied. Therefore, the manual PASI estimation may generate inconsistent scores which can mislead the judgment in Psoriasis severity (Fink et al. 2018).

Second, for dermatologists, the calculation process of PASI is time-consuming since there are 16 variables to be estiamted. In general, a professional dermatologist needs over 30 minutes to calculate PASI for a single patient (Berth-Jones et al. 2006). Considering the large number of Psoriasis patients, calculating PASI is a heavy burden for dermatologists.

Related Work

Automatic Evaluation of Psoriasis Severity

Due to the importance of automatic Psoriasis severity evaluation, extensive research efforts have been devoted to this area. Ahmand and Ihtatho (2009) first mapped skin lesions into a special-designed color space and then classified skin lesions into three different types. Lu et al. (2010) used histogram-based Bayesian classifier and support vector machine to distinguish normal skin from skin lesions and then scored erythema with K-nearest neighbor algorithm. Denmark (2004) applied Gaussian mixture model to segment skin lesions into different color channels and then scored erythema by means of the trichromatic bands.

For deep learning based approaches, Pal et al. (2016) estimated the severity levels of three indicators (the redness of erythema, the thickness of induration, and the scaling of desquamation) by adding three output heads at one single deep neural network. Pal et al. (2018) built three different networks to evaluate these three indicators respectively. In the above methods, each indicator is classified into five discrete severity categories: from 0 to 4.

Existing works mainly focus on a part of aspects related to the severity of Psoriasis. In this paper, we propose a unified

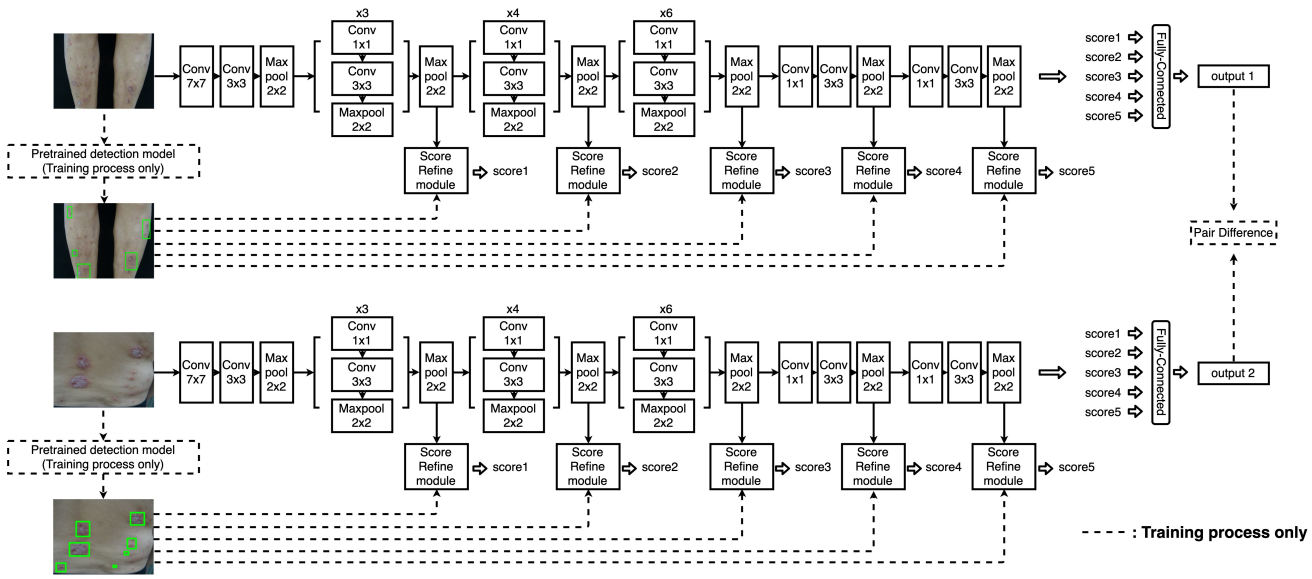

Figure 1: Structure overview of PSENet. During the training phase, each score refine module takes a location map as a supervision signal. During the testing phase, this module can generate the location map by itself.

framework to estimate the overall severity score which can be used in clinical practice directly.

Siamese Network

Siamese network has been used in various tasks to model the relationship among different inputs. This structure was first introduced by Bromley et al. (1994). Currently, many vairants of this structure have been proposed. Chopra et al. (2005) used this structure to model face verification as a metric learning problem, with the aim of discriminating different human faces. Koch, Zemel, and Salakhutdinov (2015) used a siamese-based network to project images into a low-dimensional space and evaluated the similarity among them. Zhou et al. (2019) designed a siamese network to detect salient objects by enhancing its edge. Bertinetto et al. (2016) introduced this structure into video object tracking (VOT) task. They used siamese network to extract features for each video frame and calculated the similarity between adjacent frames to track objects. Li et al. (2018) combined siamese network with region proposal network to redefine VOT task as an one-shot detection task. Moreover, siamese structure can also be applied to regression tasks. Doumanoglou et al. (2016) used siamese structure in 3D object pose estimation task. They introducd a specific loss function and used siamese structure to ensure the distribution alignment between feature space and pose space.

Method

In this section, we introduce the proposed model: PSENet. First, we introduce the framework of PSENet; secondly, we describe the specially designed score refine module in detail; and finally, we discuss the loss functions and settings.

Framework

Figure 1 displays the framework of the proposed PSENet which consists of two identical sub-networks. Each sub-network includes a backbone network which works as a feature extractor and five score refine modules which calculate the severity at varied granularities.

Inside each sub-network, we build the backbone network with 45 convolutional layers and 6 downsampling layers. Residual shortcut connections (He et al. 2016a; 2016b) are added to increase skip layer connection.

In the Psoriasis severity evaluation task, our backbone network takes images with a fixed-size: $800 * 1024$. During the propagation along the backbone network, we extract the feature maps with five resolutions: $100 * 128, 50 * 64, 25 * 32, 13 * 16$, and $7 * 8$. These feature maps are sent to score refine modules to calculate the severity. As a result, each sub-network can evaluate severity on features with different granularities. Unlike the structure proposed by Lin et al. (2017a), in which feature maps at different levels are added up together, PSENet focuses on one specific granularity at one time; thus, the unexpected interactions among feature maps and the extra upsampling/deconvolution noises can be excluded.

For each image, besides the RGB matrix, we also perform an object detection step and use the marked skin lesion area as input during the training stage. Here, we pre-trained a skin lesion detection model which is a separate model independent from the proposed PSENet. The marked skin lesions will be used to guide the training of score refine module.

In the training stage, given two input images and detected skin lesions, each sub-network generates a severity score for the corresponding input image, and the difference between two images is also obtained by using siamese structure.

In the testing stage, either sub-network can be used to

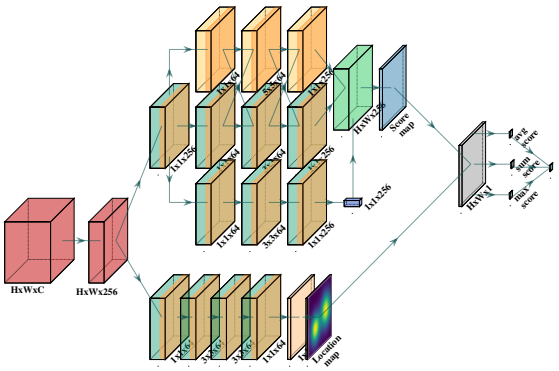

Figure 2: Structure of score refine module. The first red block denotes an input feature map with height H, width W and channel C.

evaluate the severity of an input image since the two sub-networks share all parameters and have identical structure. Besides, in the propagation within score refine module, it will generate an intermediate feature map which can locate the position of skin lesions.

Severity score labelling requires professional knowledge in dermatology and is time-consuming for dermatologists; thus, it is unlikely to obtain adequate labelled data for the model training. By using siamese structure as a training strategy, we model the severity evaluation as a metric learning task. In each training step, two images are fed into PSENet. Then, PSENet can not only learn the severity of each image, but also learn the difference between them. By introducing such a pairing scheme, we can easily generate a large number ($\sim O(N^2)$) of paired training samples by using just a small number ($\sim O(N)$) of labelled images.

Score Refine Module

Various noises exist in clinical images of Psoriasis, such as background normal skin, nevus, birthmark, and so on. These noises often occupy the most regions of the images, and therefore the skin lesions just take up few parts. To reduce the impact of noises and improve the robustness of PSENet, we design the score refine module, as illustrated in Figure 2.

This module consists of two heads, a severity evaluating head (scoring head) and a skin lesion locating head (locating head). By using the localization result as an attention mask to refine the severity evaluating result, the module can eliminate the influence of noises and focus on the skin lesion areas.

The scoring head has three sibling branches to generate pixel-wise score map O. These branches have point-wise convolution layers (Howard et al. 2017) to reduce the dimension of channel at the beginning for efficiency consideration. And then different kernel sizes have been used to add feature diversity (Szegedy et al. 2016). The first two branches are used to learn semantic information of severity and result in a feature map F, and the third branch is used to learn the weights $\vec{v}$, with the purpose of combining 256-channel feature map F into 1-channel score map O. As a result, each pixel value at the position (i, j) in score map O is calculated by weighting the 256-D vector $\vec{F}_{i,j}$ in feature map F at the corresponding position (i, j), using weight vector $\vec{v}$ from the third branch:

$$O_{i,j} = \vec{F}_{i,j} \cdot \vec{v} \tag{3}$$

To make the evaluating results more invulnerable to noise, the module locates skin lesions pixel-wisely and uses the locating result (location map D) as an attention mask to refine the scoring result. However, the prerequisite to enable a deep neural network to locate skin lesions is sufficient annotated data. Considering that there could be numerous skin lesions in a clinical image of Psoriasis, to annotate each of them in a pixel-wise fashion would be a huge burden for dermatologists. As a result, in the training process, we design the below method to generate pixel-wise ground truth maps for our locating task. We first use an external private skin disease dataset (over 86,000 images with pathologically confirmed category annotations, and location information for skin lesions in the way of bounding box provided by dermatologists) to build a simple abnormal skin detector. In this paper, the detector is built by a variant of Ren et al. (2015) with ROI-Align (He et al. 2017), and it shows the reall of 81.67% and the precision of 94.23% for Psoriasis skin lesion. Then, considering that the form of bounding box cannot match the shape of skin lesions in almost all cases, we divide the skin inside a detected bounding box into two classes, "normal" or "abnormal", based on its distance to the edge and the size of the box.

Specifically, suppose that $I \in [0, 1]^{H \times W \times C}$ is the input image with height H, width W and channel C; B is a set of all detected bounding boxes of skin lesions in I; and $M_{in} \in [0, 1]^{\frac{H}{S} \times \frac{W}{S} \times C_{M_{in}}}$ is the input feature map for a score refine module M, where S is the stride ($S \in \{8, 16, 32, 64, 128\}$). We produce ground truth map $G \in [0, 1]^{\frac{H}{S} \times \frac{W}{S}}$ for D in M.

All elements in G are generated as follows: first, for each detected bounding box $gt \in B$, we set the values at positions outside gt to 0. Second, for the center position $\vec{p}$ of gt on the input image, we project $\vec{p}$ into M_{in} as $\vec{p'} = \lfloor \frac{\vec{p}}{S} \rfloor = ({p'}_x, {p'}_y)$, and set the value at position p' to 1. Third, we use a Gaussian kernel to compute the values at the remaining positions which are inside gt using the following equation:

$$G_{i,j} = \exp\left(\frac{-(|{p'}_x - i| + |{p'}_y - j|)}{\sigma_{gt}^2}\right) \tag{4}$$

where σ_{gt} is the standard deviation for the Gaussian kernel computed based on the size of gt. Based on our experiments, we choose $\sigma_{gt} = \frac{1}{6} \times dig_{gt}$ where dig_{gt} is the projected diagonal distance of gt. Figure 3 shows an example of five ground truth maps generated by one input image during the training process.

After score map O and location map D are generated, the module then computes the element-wise product of these two maps, and obtain the refined score map. Such refining process can be seen as a hard-supervised attention mechanism.

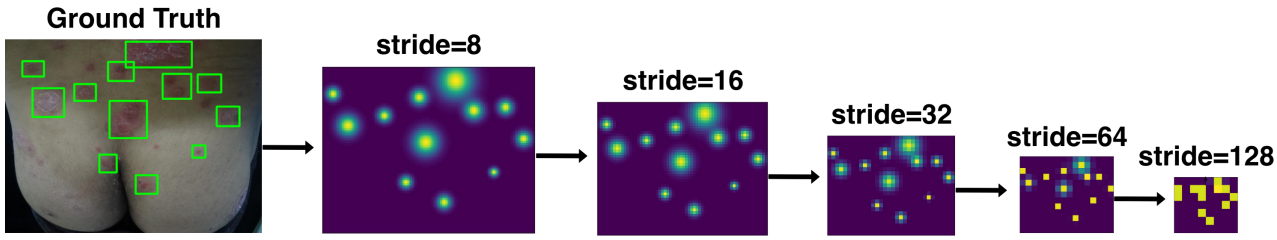

Figure 3: An example of five ground truth maps generated by an input image during the training process. The locations of skin lesions are generated by Ren et al. (2015) with ROI-Align (He et al. 2017). In ground truth maps, the color from purple to yellow denotes the value of element from 0 to 1, representing the confidence of the skin lesion existence.

From a dermatologist's point of view, the general severity and the severity of targeting skin lesion (the most severe one among all skin lesions) are two key factors for evaluating. As a result, we extract three different statistics from the refined score map (max score, mean score, and sum score) and combine them together by three trainable weights to get the final score. All parameters in five score refine modules are shared except for these three weights, meaning that all five score refine modules have exactly identical branches (scoring, locating, and refining) except for the final combination weights. This is because the three statistics from different features have different distributions. In fact, these three weights play a role in regulating the contributions of three statistics in a single module and balancing the scales among the outputs of different modules.

Loss Functions

We build an end-to-end framework for the task of Psoriasis severity evaluation, which includes two sub-tasks for the skin lesion locating and distance metric learning, as well as the main task for severity score regression. Next, we introduce the loss function for each task and illustrate the overall loss function.

Localization loss Each score refine module locates skin lesions in a pixel-wise fashion. Since the background always takes up the most area of an image, it is necessary to balance the proportion between positive and negative samples for this task. We design the loss function based on focal loss (Lin et al. 2017b). It not only balances the proportion between positive and negative samples, but also makes sure that the gradient is not dominated by non-hard-samples. The localization loss is designed as follows:

$$l_{i,j} = \begin{cases} -(1-\alpha)(1-D_{i,j})^{\gamma} log D_{i,j}, & G_{i,j} \geq thr \\ -\alpha D_{i,j}^{\gamma} log(1-D_{i,j}), & G_{i,j} < thr \end{cases} \tag{5}$$

$$L_{loc} = \sum_{k=1}^{5} \frac{1}{n_k} \sum_{i,j} l_{i,j} \tag{6}$$

where n_k denotes the number of points in the point set $\{(0,0),...,(\frac{H}{S_k},\frac{W}{S_k})\}$; α and γ are the parameters defined as same as in Lin et al. (2017b); $D_{i,j}$ denotes the value of the position (i,j) in the location map; $G_{i,j}$ is the value of the position (i,j) in the ground truth map; and thr is the threshold to divide positive and negative samples for elements on the ground truth map. We use $\alpha = 0.25$, $\gamma = 2$, $p = 5$, and thr = 0.3 in our experiments.

Distance metric loss To learn the metric of severity and the discrepancy between a pair of images, we introduce siamese structure into our framework (two sub-networks are named as network A and network B, respectively). During the training process, we use $t = \{(x_A, y_A), (x_B, y_B)\}$ as the input sample where x_A/x_B is an image in network A/B and y_A/y_B is its severity score annotation given by dermatologists. Network A predicts score s_A for x_A, and network B predicts score s_B for x_B. Next, siamese structure uses $\Delta_p = f(s_A, s_B)$ as the output and $\Delta_a = f(y_A, y_B)$ as annotation to calculate the siamese loss (f is a distance measurement, in which Manhattan distance is used in this paper). We design the final distance metric loss function based on smooth L1 loss to reduce the impact of outliers and maintain the stability of training stage (Ren et al. 2015):

$$L_{sia} = \begin{cases} 0.5 \times (\Delta_p - \Delta_a)^2 & (\Delta_p - \Delta_a) < 1 \\ |\Delta_p - \Delta_a| - 0.5 & otherwise \end{cases} \tag{7}$$

Regression loss In our model, each sub-network in siamese structure uses a fully connected layer to combine the scores from the five score refine modules into an overall score. Similar to metric loss, because of the wild range of PASI, we use smooth L1 loss for severity score regression task to prevent the oscillation in the initial training stage. For $t = \{(x_A, y_A), (x_B, y_B)\}$ and $\{s_A, s_B\}$, the regression loss is defined as:

$$L_{reg} = \begin{cases} \sum_{i}^{A,B} 0.5 \times (s_i - y_i)^2, & (s_i - y_i) < 1 \\ \sum_{i}^{A,B} |s_i - y_i| - 0.5, & otherwise \end{cases} \tag{8}$$

Overall loss function With these losses defined as above, the overall loss function for t is defined as:

$$L = \mu \times L_{loc} + \nu \times L_{sia} + \xi \times L_{reg} \tag{9}$$

where $\mu = 1.0, \nu = 0.2$, and $\xi = 0.2$ based on our experiments.

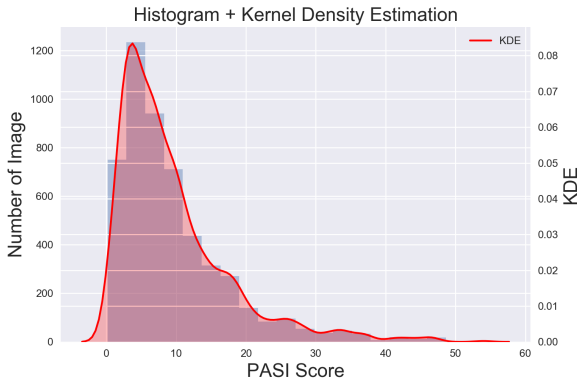

Figure 4: Severity score distribution of the dataset we built. The range of the annotated scores is from 0 to 72. As same as the situation in real life, in this dataset, the majority of severity is mild and medium (dermatologists tend to use 5 and 10 as the thresholds for mild-medium and medium-severe empirically).

Experiments

Dataset and Setting

Dataset In this study, we tracked the entire treatment processes of 1,787 Psoriasis patients and built a dataset consisting of 5,205 images where the longest recorded period is 15 months with 6 visits. The labels in this dataset include the severity scores which are annotated by 11 professional dermatologists (9 professors and 2 attending physicians), and the locations of skin lesions which are generated by our pre-trained detection model. The severity score distribution is shown in Figure 4.

Training pairs construction We split our data into 5 folds using individual patient as the smallest unit. The model is trained on 4 folds and evaluated on the held-out fold. All reported results are the average on 5 different validation folds. Paired images are randomly sampled at the beginning of each training step. Each image is resized to height with 800 and width with 1024.

Implementation The model structure and hyper-parameters in loss functions have been shown in the above sections. The model is initialized with weights pre-trained on ImageNet (Deng et al. 2009), and Adam optimizer (Kingma and Ba 2014) has been used where the initial learning rate is 0.0001 and the weight decay is 0.0002.

Baseline Comparison

We chose two representative structures ResNet-50 and GoogLeNet-v2 as our baselines, and used the regression loss in our model to be their cost functions. To be fair, we also conducted experiments of using siamese structure on these two networks and added corresponding loss to their cost functions (Table 1).

Figure 5 and Figure 6 show the statistics of the MAE of the baselines and our method. These results demonstrate that our method could generate reasonable scores in most cases, and it clearly outperforms the other two methods.

However, as reflected by Figure 7, the extreme cases whose score range is bigger than 20, are obviously difficult to evaluate. Excluding these extreme samples, the MAE of our method could decrease from 2.21 to 1.98. A possible reason for this situation is the low occurrence rate of extremely severe cases in the real world, and thus the data of such samples is insufficient.

Table 1: Performance Comparison with Baselines

Method	w/wo Sia	MAE	PA (%)
ResNet-50	no	3.50	69.36
ResNet-50	yes	3.30	70.08
GoogLeNet-v2	no	3.92	70.36
GoogLeNet-v2	yes	3.25	70.72
PSENet(ours)		**2.21**	**77.87**

w/wo denotes with/without.

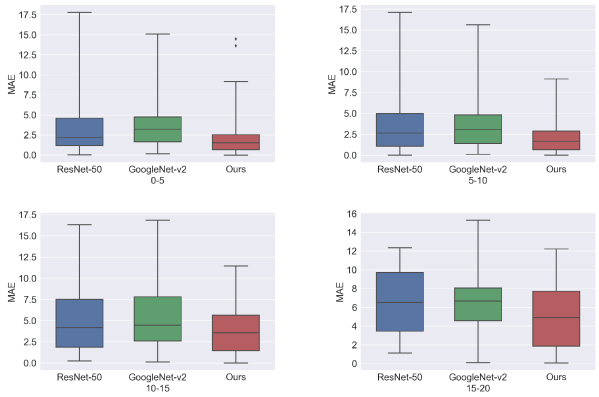

Figure 5: Comparison of the box-plot chart in terms of average MAE of five validation sets. We separately calculate the Median, Quartile, and Extreme Points for samples in different score ranges: 0-5, 5-10, 10-15 and 15-20.

Ablation Study

The main idea of our method lies in score refine module and siamese structure. Next, we made ablation studies on each of them.

Table 2: Ablation Study Results

Model	fp/srm	w/wo Sia	MAE	PA (%)
PSENet	fp	no	3.27	71.32
PSENet	srm	no	2.76	76.96
PSENet	fp	yes	3.03	73.76
PSENet	srm	yes	**2.21**	**77.87**

fp denotes feature pyramid;
srm denotes score refine module.

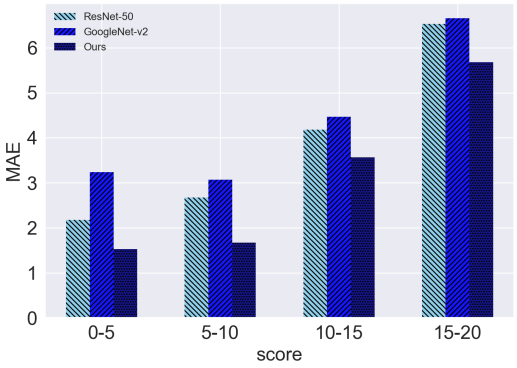

Figure 6: Comparison of the bar chart in terms of average MAE of five validation sets. We separately calculate the MAE for samples in different score ranges: 0-5, 5-10, 10-15 and 15-20.

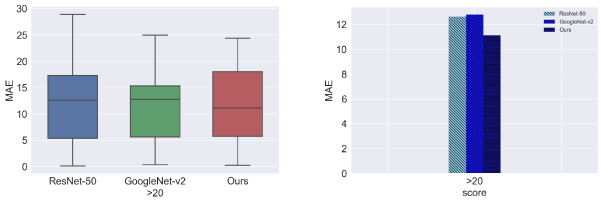

Figure 7: Comparison of the box plot char and bar chart in terms of MAE for samples with annotation score larger than 20.

Score refine module To evaluate its validity, we compared the results of our model with and without score refine modules. In experiments without score refine modules, we directly used a global average pooling on five feature maps in the feature pyramid to generate severity scores. Finally, we used the same strategy to combine the resulting five scores to generate the overall severity score. From the results shown in Table 2, we can see that the method with score refine moduless perform better than its competitor. For example, without siamese structure, the MAE of the model with score refine modules improves by 0.51 against its competitor, and the PA of the model with score refine modules increases by 5.64% against its competitor. With siamese structure, the MAE and the PA of the model with score refine modules improve by 0.82 and 4.11% against its competitor, respectively.

The performance gain comes from that the module uses a skin lesion locating result as attention mask to alleviate the impact from various noises, which makes the model focus on skin lesions truely useful for severity evaluation task. We visualized the attention mask in five score refine modules in Figure 8. From the visualization results, we observe that the module indeed pays more attention to the area with skin lesions, meanwhile ignores most of background and normal skin areas.

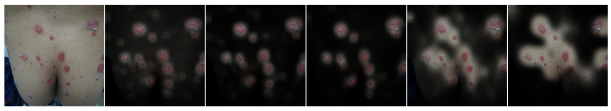

Figure 8: Localization results generated by five score refine modules. The original sizes of these results are $100*128$, $50*64$, $25*32$, $13*16$, and $7*8$, respectively. We resized them to the resolution as same as the input images by Bicubic Interpolation for better presentation. The darkness denotes the element value from 0 to 1.

Siamese structure We used siamese network to handle the problem of data insufficiency. Table 1 and Table 2 compared the performance of different models with and without siamese structure. In Table 2, without score refine module, the model using siamese structure outperforms its competitor: the MAE improves by 0.24 and the PA increases by 2.44%. With score refine module, the model using this structure improves the performance by 0.55 in terms of the MAE and 0.91% in terms of the PA. These results clearly justify the effectiveness of siamese structure in our model.

Conclusions

Quantitatively measuring the severity of Psoriasis is a crucial task. The increase or decrease of severity can indicate the progress of Psoriasis. If the score increases, which means that the condition of a patient is getting worse, the dermatologist needs to take timely action, such as adjusting current therapy; otherwise, it could cause serious complications. Currently, PASI is the most frequently used metric to evaluate Psoriasis severity. PASI is manually calculated by dermatologists which is time-consuming. Furthermore, the calculation of PASI includes 16 variables to be estimated which is error-prone and can easily bring in variations. For the same patient, different dermatologists may come up with different results. Such inconsistency makes it hard to track Psoriasis progress. To overcome these drawbacks, we propose an automatic method, PSENet. PSENet is an end-to-end framework which generates a numeric severity score for an input clinical image. PSENet introduces a specially designed module, score refine module, to localize and evaluate severity of skin lesion. Furthermore, it uses siamese structure as a training strategy to learn the difference between a pair of images; thus reducing the dependency on larger amount of labeled images. PSENet achieves the MAE of 2.21 and the PA of 77.87%, outperforming baseline models. Using this method, dermatologists could be relieved from the repetitive and dull calculation process of PASI, and Psoriasis patients could track their severity in a much more convenient way.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China under Grant No.2018YFC0117000, in part by the National Natural Science Foundation of China under Grant No.81573049, and in part by the Natural Science Foundation of Hunan Province of China under Grant No.2018JJ3689.

References

Ahmand, M., and Ihtatho, D. 2009. Objective assessment of psoriasis erythema for pasi scoring. *J. Med. Eng. Technol*.

Berth-Jones, J.; Grotzinger, K.; Rainville, C.; Pham, B.; Huang, J.; Daly, S.; Herdman, M.; Firth, P.; and Hotchkiss, K. 2006. A study examining inter-and intrarater reliability of three scales for measuring severity of psoriasis: Psoriasis area and severity index, physician's global assessment and lattice system physician's global assessment. *British Journal of Dermatology* 155(4):707–713.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, 850–865. Springer.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, 737–744.

Chopra, S.; Hadsell, R.; LeCun, Y.; et al. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, 539–546.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Denmark, D. D. 2004. An image based system to automatically and objectively score the degree of redness and scaling in psoriasis lesions. In *Proceedings fra den 13. Danske Konference i*, 130.

Doumanoglou, A.; Balntas, V.; Kouskouridas, R.; and Kim, T.-K. 2016. Siamese regression networks with efficient mid-level feature extraction for 3d object pose estimation. *arXiv preprint arXiv:1607.02257*.

Fink, C.; Fuchs, T.; Enk, A.; and Haenssle, H. A. 2018. Design of an algorithm for automated, computer-guided pasi measurements by digital image analysis. *Journal of Medical Systems*.

George, Y.; Aldeen, M.; and Garnavi, R. 2017. Automatic psoriasis lesion segmentation in two-dimensional skin images using multiscale superpixel clustering. *Journal of Medical Imaging*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645. Springer.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Howard, A. G.; Zhu, M.; Bo, C.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Langley, R. G., and Ellis, C. N. 2004. Evaluating psoriasis with psoriasis area and severity index, psoriasis global assessment, and lattice system physician's global assessment. *Journal of the American Academy of Dermatology* 51(4):563–569.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8971–8980.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lu, J.; Manton, J. H.; Kazmierczak, E.; and Sinclair, R. 2010. Erythema detection in digital skin images. In *2010 IEEE International Conference on Image Processing*, 2545–2548. IEEE.

Pal, A.; Chaturvedi, A.; Garain, U.; Chandra, A.; and Chatterjee, R. 2016. Severity grading of psoriatic plaques using deep cnn based multi-task learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1478–1483. IEEE.

Pal, A.; Chaturvedi, A.; Garain, U.; Chandra, A.; Chatterjee, R.; and Senapati, S. 2018. Severity assessment of psoriatic plaques using deep cnn based ordinal classification. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer. 252–259.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Zhou, Y.; Sheng, Y.; Gao, J.; and Zhang, X. 2015. Dermatology in china. In *Journal of Investigative Dermatology Symposium Proceedings*, volume 17, 12–14. Elsevier.

Zhou, S.; Wang, J.; Wang, F.; and Huang, D. 2019. Se2net: Siamese edge-enhancement network for salient object detection. *arXiv preprint arXiv:1904.00048*.