Joint Acne Image Grading and Counting via Label Distribution Learning

Xiaoping Wu[1*], Ni Wen[2*], Jie Liang[1], Yu-Kun Lai[3], Dongyu She[1], Ming-Ming Cheng[1], Jufeng Yang[1✉]
[1]College of Computer Science, Nankai University [2]Beijing Tsinghua Changgung Hospital
[3]School of Computer Science and Informatics, Cardiff University
xpwu95@163.com, nini1992713@126.com, liang27jie@163.com, LaiY4@cardiff.ac.uk,
sherry6656@163.com, {cmm, yangjufeng}@nankai.edu.cn

Abstract

Accurate grading of skin disease severity plays a crucial role in precise treatment for patients. Acne vulgaris, the most common skin disease in adolescence, can be graded by evidence-based lesion counting as well as experience-based global estimation in the medical field. However, due to the appearance similarity of acne with close severity, it is challenging to count and grade acne accurately. In this paper, we address the problem of acne image analysis via Label Distribution Learning (LDL) considering the ambiguous information among acne severity. Based on the professional grading criterion, we generate two acne label distributions considering the relationship between the similar number of lesions and severity of acne, respectively. We also propose a unified framework for joint acne image grading and counting, which is optimized by the multi-task learning loss. In addition, we further build the ACNE04 *dataset with annotations of acne severity and lesion number of each image for evaluation. Experiments demonstrate that our proposed framework performs favorably against state-of-the-art methods. We make the code and dataset publicly available at* `https://github.com/xpwu95/ldl`*.*

1. Introduction

Automatic grading of skin disease severity is of great importance in the medical field. Acne vulgaris, commonly named acne, is the most common skin disease, which has prevalence peaks during adolescence [28, 33]. About 80% adolescents suffer from acne [9] and the symptoms last through adulthood in 3% men and 12% women [23]. Therefore, there are a massive number of acne patients that need specific treatment imminently, since acne may also leave scars and pigmentation and often leads to considerable inferior and depressed emotions [48]. The severity of acne is essential for dermatologists to make a precise and standardized treatment decision [24]. Besides, junior dermatologists also need an objective and reliable diagnosis for reference. A standard criterion used by dermatologists for grading acne severity is the Hayashi criterion [24], which combines the outcome measures of lesion counting and global assessment. Specifically, acne can be graded into four levels of severity, *i.e*. mild, moderate, severe, and very severe, according to the number of lesions.

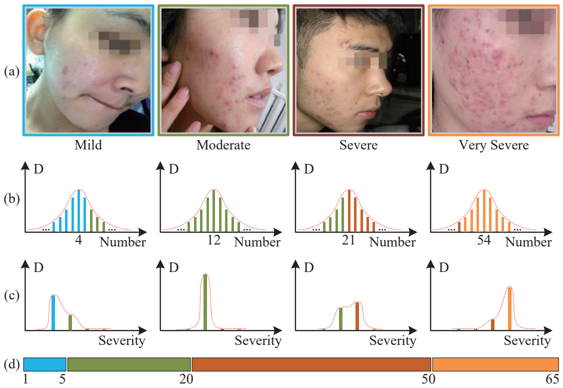

Figure 1. Image examples (a) and their corresponding label distributions (b, c). The x-axes in (b, c) indicate the number of lesions and acne severity, respectively. The label distributions cover several neighboring labels which represent the degree each label describes the instance (denoted as "D"). Acne images can be divided into different severity levels according to the lesion numbers [24] (d). Different colors indicate different severity levels, *e.g*., blue is for mild, and green is for moderate.

In the past years, substantial advances have been made for acne lesion analysis [2, 6, 16]. Most methods indirectly focus on the classification or detection of acne lesions and generally rely on hand-crafted features. For example, Abas *et al*. [1] employ discrete wavelet frames and gray-level co-occurrence matrices to extract features for detecting acne lesions. Recently, Convolutional Neural Net-

[*]Equal contributions

works (CNNs) [39, 35, 56] show powerful performance in the medical imaging processing tasks, *e.g*., common thorax disease classification [47] and biomedical segmentation [8]. However, there are some limitations when employing CNNs for acne image analysis. First, acne images with close severity show similar appearance, while the existing single-label learning methods (SLL) [25] represent the acne label using one-hot vector ignoring the ambiguity issue, as shown in Fig. 1(a). Second, the tasks of lesion counting and acne grading have different objectives, *i.e*., the classification score and counting number, which cannot be combined directly for grading acne severity.

In this paper, we address acne image analysis via Label Distribution Learning (LDL) [22], which assigns each instance with a label distribution containing the degree each label describes it. Instead of using a single label for an acne image, we propose two acne label distributions to represent the lesion number and acne severity, respectively. As illustrated in Fig. 1(b), label distribution for the lesion number is generated based on the Gaussian distribution, where the original dominant label keeps the highest description degree, and labels far from it have a lower degree. For the acne severity, since acne images belonging to the same severity level may have a greatly varied number of lesions [24], we consider the professional medical criterion to generate distribution, as shown in Fig. 1(c). We further propose a unified deep framework with two branches for joint acne severity grading and lesion counting. The counting branch first predicts the label distribution for the lesions, which is then mapped to the acne severity distribution based on the Hayashi criterion [24]. The grading branch combines the predicted severity distribution and the mapped distribution for acne image grading. Our framework is then optimized by the multi-task learning loss through end-to-end training.

Our contributions are three-fold: First, oriented by accepted medical criterion, we present a unified acne severity grading framework, which considers the procedures of global acne assessment and lesion counting simultaneously for acne image analysis. Second, we generate two acne label distributions based on the professional grading criterion, considering the relationship between the similar number of lesions and severity of acne, respectively. Third, we collect a new dataset *ACNE04*, which provides the annotations of acne severity and the bounding boxes of lesions annotated by professional dermatologists. The experimental results demonstrate that the proposed method performs favorably against the state-of-the-art methods.

2. Related Work

2.1. Medical Disease Diagnosis

Medical disease diagnosis attracts more and more attention of researchers in the vision community. Deep learning technologies which achieve significant performance on many computer vision tasks (*e.g*., classification [25, 19], detection [15, 14, 58], and segmentation [12, 13]) have been successfully employed in the medical field. Focusing on the task of medical image diagnosis, [10] utilizes deep CNNs to diagnose skin cancers from dermoscopic images. Wang *et al*. [47] jointly train a CNN-RNN model and achieve multi-label classification and reporting of common thorax diseases. Their experimental results show excellent feature representation ability of deep networks for medical images.

There are massive datasets for common object recognition [36, 11] in the vision community. However, it is challenging and expensive to collect and annotate medical images due to the requirements of expert knowledge and medical experience. Recently, [41] proposes a benchmark dataset, named SD-198, for common skin disease recognition on clinical images. Wang *et al*. [46] present the ChestX-ray8 with weakly-supervised annotations for classifying and localizing common thorax diseases on X-ray imaging. Several related works also demonstrate the importance of professional medical criteria for medical disease diagnosis. Yang *et al*. [53] design a computer-aided diagnosis system which represents skin lesions with several medical representations according to different criteria and achieve comparable results with dermatologists.

2.2. Object Counting

Object counting technologies are widely applied in a variety of scenarios, such as crowd counting [5, 56] and vehicle counting [32]. They can be mainly grouped into two types: detection and regression based methods. Detection-based mechanisms aim to detect the specific location and size of objects and then convert the proposals into the counting result. [44, 59] are first used to generate potential object proposals. Then the classifiers are trained with hand-crafted features [7, 37] or recent deep features [35]. The proposals with high classification confidence are counted for the final result. Besides, state-of-the-art object detection methods [55, 34] seek the end-to-end training architecture for real-time applications. Detection based methods perform well to some degree, but there are still challenges, *e.g*. detected objects are too small in most counting scenarios. Without concerning the detailed location of objects, regression-based methods globally estimate counting results [29] from features. To maintain spatial information, [3, 4] consider regressing the density map from the feature map of CNNs and achieve more accurate performance. The large-scale variation also cannot be ignored in real life applications. [38, 56] design scale-aware networks to adapt to the density variation in input images. Meanwhile, describing the inherent characteristic that counted images with neighboring class labels have similar features [57] is of importance to regression-based methods.

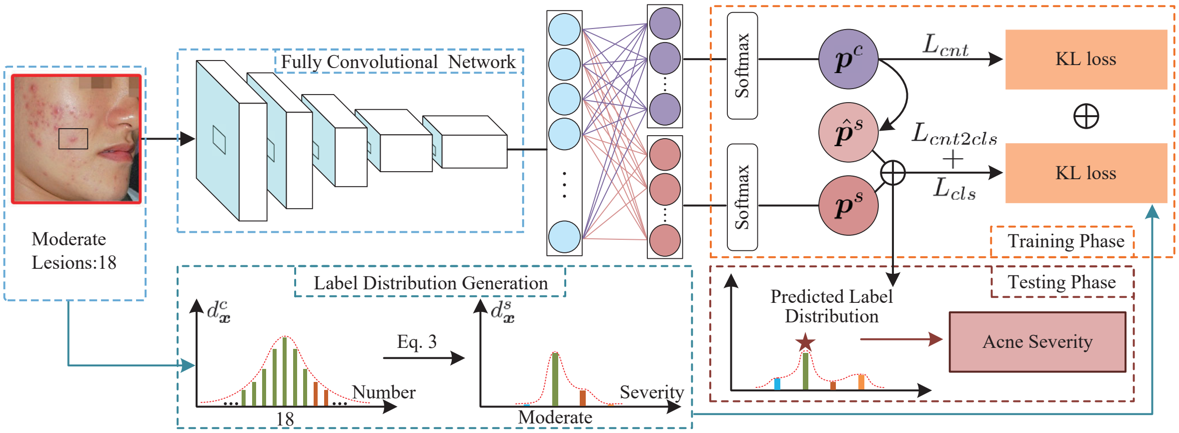

Figure 2. Pipeline of the proposed framework. The input image is resized and passed through the CNN backbone model (ResNet-50 [25]). Then the framework is divided into two branches. The grading branch globally estimates the severity of acne. The counting branch first predicts the label distribution of acne lesion count. Then it is converted into the label distribution of acne severity based on Eq. 3. The counting model simultaneously grades the acne severity and predicts the lesion number to provide acne diagnosis evidence. Finally, the prediction results from global grading and local counting models are merged, oriented by the medical criterion.

2.3. Label Distribution Learning

To cope with the issue of label ambiguity existing in traditional single label learning, Geng *et al*. [22, 45] propose a new machine learning paradigm, *i.e*., label distribution learning. Instead of assigning a single label [25], LDL covers a certain number of adjacent labels where each label represents different description degree to the instance, respectively. Recently, LDL has been used to well address label ambiguities in various tasks, *e.g*., age estimation [26, 27, 54], head pose estimation [21], and visual sentiment analysis [51, 52]. Many methods [49, 40, 18, 50] successfully employ the Gaussian function to generate label distribution. Geng *et al*. [20] propose adaptive label distribution learning (ALDL) to generate label distributions with different shapes for each age, *i.e*., different variance parameter in the Gaussian function for each class. In this particular scenario that the aging process varies at different aging stages, the ALDL performs well with sufficient labeled training data. Subsequently, Hou *et al*. [27] propose a semi-supervised ALDL method that utilizes the unlabeled data to solve the data-scale problem. In addition, Gao *et al*. [17] employ deep CNNs for label distribution learning by minimizing the KL divergence between the predicted and ground-truth distributions. In the multi-task learning scenario, our framework also explores the latent fusion of multiple tasks in both training and testing phases.

3. Method

Fig. 2 illustrates the pipeline of our proposed method with two branches for joint acne severity grading and lesion counting. Given N input training images with corresponding single labels of acne severity and ground-truths of lesion counts $\{(\boldsymbol{x}_1, y_1, z_1), \cdots, (\boldsymbol{x}_N, y_N, z_N)\}$, where $y_i \in [1, \cdots, Y]$ and $z_i \in [1, \cdots, Z]$. Y and Z denote the class number of acne severity levels and max number of lesion counts, respectively. The goal of our framework is to simultaneously output the grading result of acne severity and the counting result of lesions as diagnostic evidence. The following subsections introduce details of these two tasks respectively and the final multi-task learning strategy.

3.1. Label Distribution Generation

For the input image $\boldsymbol{x}_i$, we utilize the Gaussian function following [17] to generate the label distribution for the lesion counting task. The description degree of one particular count label of acne lesions c_j to the instance $\boldsymbol{x}_i$ can be defined as:

$$d^{c_j}_{\boldsymbol{x}_i} = \frac{1}{\sqrt{2\pi}\sigma M} \exp\left(-\frac{(c_j - z_i)^2}{2\sigma^2}\right), \qquad (1)$$

where $j \in [1, \cdots, Z]$ and all the labels are utilized to describe the instance. The standard deviation σ is a hyperparameter which controls the distribution amplitude, which is set as 3 in this paper. Let the vector $\boldsymbol{d}^c_{\boldsymbol{x}_i} = [d^{c_1}_{\boldsymbol{x}_i}, \cdots, d^{c_Z}_{\boldsymbol{x}_i}]$ denote the label distribution of the instance $\boldsymbol{x}_i$ in the counting task. The label distribution $\boldsymbol{d}^c_{\boldsymbol{x}_i}$ generated by a Gaussian function has two properties. The first is that $d^{c_j}_{\boldsymbol{x}_i} \in [0, 1]$ and $\sum_{j=1}^{Z} d^{c_j}_{\boldsymbol{x}_i} = 1$. The normalization factor M ensures this property, where

$$M = \frac{1}{\sqrt{2\pi}\sigma} \sum_{j=1}^{Z} \exp\left(-\frac{(c_j - z_i)^2}{2\sigma^2}\right). \qquad (2)$$

The other is that the ground-truth label of lesion count describes the highest degree, *i.e*., the description degree

$d_{\boldsymbol{x}_i}^{z_i} \geqslant d_{\boldsymbol{x}}^{c_j}$. The count label farther from the ground-truth has lower description degree.

In the grading task, the label distribution is converted from the counting task. Specifically, Table 1 shows that the count of acne lesions in an image can be mapped to a specific class of acne severity according to the medical criterion [24]. Then the description degree $d_{\boldsymbol{x}_i}^{s_k}$ of the acne severity label to instance $\boldsymbol{x}_i$ can be defined as the sum of the description degrees of the lesion count labels that belong to the corresponding mapping interval $\phi(k)$ based on the medical criterion:

$$d_{\boldsymbol{x}_i}^{s_k} = \sum_{j \in \phi(k)} d_{\boldsymbol{x}_i}^{c_j}, \tag{3}$$

where $k \in [1, \cdots, Y]$. The label distribution $\boldsymbol{d}_{\boldsymbol{x}_i}^{s} = [d_{\boldsymbol{x}_i}^{s_1}, \cdots, d_{\boldsymbol{x}_i}^{s_Y}]$ used for the grading task also satisfies the aforementioned two properties similar to the counting task.

3.2. Lesion Counting

For the input instance $\boldsymbol{x}_i$, its predicted probability of belonging to each class $j \in \{1, \cdots, Z\}$ is calculated as:

$$p_i^{(j)} = \frac{\exp(\theta_j)}{\sum_{m=1}^{Z} \exp(\theta_m)}, \tag{4}$$

where θ_j is the predicted score corresponding to the j-th class outputted from the last fully connected layer. We apply the KL loss following [17] to minimize the deviation between ground-truth label distribution $\boldsymbol{d}_{\boldsymbol{x}_i}^{c}$ and predicted label distribution $\boldsymbol{p}_i^c = [p_i^{(1)}, \cdots, p_i^{(Z)}]$ in the counting task:

$$L_{cnt}(\boldsymbol{x}_i, z_i) = -\sum_{j=1}^{Z} \left(d_{\boldsymbol{x}_i}^{c_j} \ln p_i^{(j)} \right). \tag{5}$$

Observed from the Hayashi criterion [24], we can find that counting has certain practical significance in the task of acne severity grading. The information of lesion count can latently classify the acne into one of the four severity levels (*i.e.*, mild, moderate, severe and very severe) according to the mapping intervals $\phi(k)$ as shown in Table 1. Hence we further convert the predicted counting result $\boldsymbol{p}_i^{(c)}$ to grading result $\hat{\boldsymbol{p}}_i^s = [\sum_{j \in \phi(1)} p_i^{(j)}, \cdots, \sum_{j \in \phi(Y)} p_i^{(j)}]$ based on Eq. 3. Then the loss of label distribution of grading results converted from the counting results can be defined as:

$$L_{cnt2cls}(\boldsymbol{x}_i, y_i) = -\sum_{k=1}^{Y} \left(d_{\boldsymbol{x}_i}^{s_k} \ln \sum_{j \in \phi(k)} p_i^{(j)} \right). \tag{6}$$

3.3. Acne Severity Grading

The previous section shows that the counting task can provide the results of acne severity and lesion number simultaneously. Yet the procedure of global acne severity grading is also necessary, since the Hayashi criterion [24] grades the acne severity based on the combination of global and local diagnosis results.

For the i-th input instance $\boldsymbol{x}_i$, its predicted probability of belonging to each class $k \in \{1, \cdots, Y\}$ is calculated as:

$$p_i^{(k)} = \frac{\exp(\delta_k)}{\sum_{n=1}^{Y} \exp(\delta_n)}, \tag{7}$$

where δ_k is the predicted score corresponding to the k-th class outputted from the last fully connected layer. The distribution loss of $\boldsymbol{p}_i^s = [p_i^{(1)}, \cdots, p_i^{(Y)}]$ in the KL divergence form can be defined as:

$$L_{cls}(\boldsymbol{x}_i, y_i) = -\sum_{k=1}^{Y} \left(d_{\boldsymbol{x}_i}^{s_k} \ln p_i^{(k)} \right). \tag{8}$$

3.4. Multi-Task Learning Model

Different losses or tasks mentioned above guide the model to focus on different aspects of acne images. For example, the classification loss makes global estimation and the counting loss tends to explore local information of specific lesions. A unified multi-task learning strategy latently leads the model to learn more robust and discriminative description of features and classifier.

Our model combines the advantages of both global and local features for visual acne representations both in training and testing phases. At the training procedure, the multi-task learning loss is defined as:

$$\begin{aligned} L_i(\boldsymbol{x}_i, y_i, z_i) =& (1-\lambda) L_{cnt}(\boldsymbol{x}_i, z_i) \\ &+ \frac{\lambda}{2} \left(L_{cls}(\boldsymbol{x}_i, y_i) + L_{cnt2cls}(\boldsymbol{x}_i, y_i) \right), \end{aligned} \tag{9}$$

where the hyper-parameters of λ is the trade-off between counting and grading tasks.

At the testing phase, the model merges classification results from grading task $\boldsymbol{p}_i^y$ and counting task $\hat{\boldsymbol{p}}_i^y$ for the instance $\boldsymbol{x}_i$. The final diagnosis takes the average of them $\frac{1}{2}(\boldsymbol{p}_i^y + \hat{\boldsymbol{p}}_i^y)$. In this way, our method achieves an end-to-end procedure that simultaneously grades the acne severity and provides diagnostic evidence of lesion counts. Besides, it combines the global estimation and lesion counting tasks both in training and testing phases.

4. Experiments

In this section, we detail the experiment settings, parameters, ablation analysis, and comparison with the state-of-the-art methods.

4.1. Dataset & Evaluation Metrics

For verifying our algorithm and promoting further study on medical disease grading, we build an acne severity grading dataset named *ACNE04*. The *ACNE04* dataset includes the annotations of local lesion numbers and global acne severity. When the experts are making a diagnosis, the images with acne lesions are collected by a digital camera with the consent of patients. Following the requirements of the Hayashi grading criterion [24], all images are taken at an approximately 70-degree angle from the front of patients. Then the experts manually annotate the images with our provided annotation tool. Fig. 3 shows several example images with annotations. Under the challenges both on the procedures of data collection and annotation, the *ACNE04* contains $1,457$ images with $18,983$ bounding boxes of lesions. For evaluating, we split the dataset into 80% training set and 20% testing set, containing $1,165$ and 292 images, respectively, as shown in Table 1.

Following previous methods, we select different evaluation metrics for the tasks of classification and object counting, respectively. The commonly utilized accuracy and precision are applied to evaluate the classification performance. Considering that our work of the acne severity grading is related to medical image processing, we additionally choose several important metrics from the medical field, including sensitivity, specificity, and Youden Index. Sensitivity is typically named recall or true positive rate in the vision community. Specificity is the true negative rate, which reflects the ability to correctly rule out a disease. Youden Index equals $(\text{Sensitivity} + \text{Specificity} - 1)$ with a range of $[-1, 1]$, which represents the comprehensive ability of diagnosis. Larger Youden Index indicates higher diagnostic value and the diagnosis is entirely meaningless if it is less than or equal to 0. We adopt mean absolute error (MAE) and root mean squared error (MSE) to evaluate the object counting performance [56, 4].

4.2. Implementation Details

The backbone of our architecture is ResNet-50 [25] with the parameters pre-trained on the ImageNet [36] dataset. Before training the network, we resize the input image to $224 \times 224 \times 3$ pixels and normalize it to the range of $[0, 1]$ in RGB channels, respectively. We choose Stochastic Gradient Descent (SGD) with the mini-batch of 32 as the model optimizer and train the model for 120 epochs ensuring that the average loss on the training set is stable. The momentum and weight decay are set to 0.9 and $5e$-4, respectively. We start the learning rate at 0.001 and decay it by 0.5 every 30 epochs. Our algorithm runs on an NVIDIA TITAN X GPU with 12GB VRAM. Notice that 5-fold cross-validation is applied for robust evaluation. Our proposed algorithm is implemented based on the PyTorch framework.

4.3. Parameters

In this section, we experimentally discuss the setting of λ parameter. The λ parameter is the trade-off between the acne grading and lesion counting tasks. Larger λ makes the model pays more attention on the counting task. We evaluate the proposed algorithm performance under different settings of λ from 0.1 to 0.9 using the accuracy and MAE metrics. As illustrated in Fig. 4, with an increasing λ, the model performs better within a certain range. The model gains the best performances on the two evaluation metrics when $\lambda = 0.6$. So we choose $\lambda = 0.6$ as the final parameter setting.

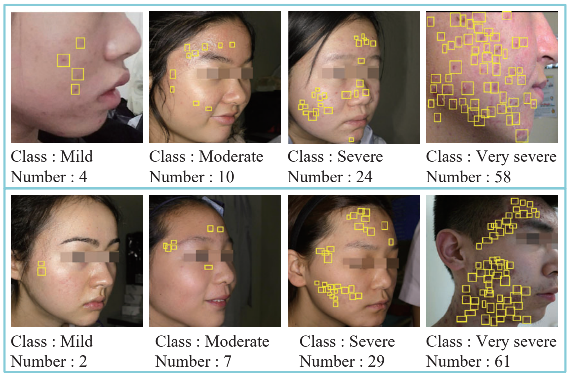

Figure 3. Examples in the *ACNE04* dataset. The numbers under each image denote the ground-truth severity and lesion number, respectively. The yellow bounding boxes represent the lesion positions.

Table 1. Medical criterion [24] and statistics of training and testing splits of the *ACNE04* dataset. The criterion represents the relationship between severity class and lesion number.

Class	Criterion	Training		Testing	
		Image	Lesion	Image	Lesion
Mild	$1 \sim 5$	410	858	103	221
Moderate	$6 \sim 20$	506	4,547	127	1,123
Severe	$21 \sim 50$	146	3,857	36	890
Very severe	> 50	103	5,965	26	1,522
Total	-	1,165	15,227	292	3,756

4.4. Ablation Studies

In this section, we analyze the efficiency of each component in our proposed method. The conventional SLL utilizes a single label to represent the instance. Table 2 shows that this learning scheme achieves the accuracy of 78.42% on the acne grading task and the MAE of 4.16 on the lesion counting task. The counting results are converted into corresponding acne severity guided by the Hayashi criterion [24] and achieves the accuracy of 75.69%.

We first introduce the LDL into the grading and counting tasks, respectively. On the grading task, the LDL gains slightly with improved accuracy of 0.89% compared with the SLL. Yet the very low standard deviation (less than 0.1)

indicates that it can achieve more reliable acne grading results. On the counting task, the LDL improves the model performance in the MAE by 0.92. In addition, the converted grading results gain significant improvements on all the evaluation metrics, *e.g.*, by 5.40% for the accuracy, even better than the direct grading procedure. This indicates that the label distribution of the lesion number has the latent capacity to represent the continuous features of acne images. The improvements also demonstrate that it is reasonable and capable to discriminate acne severity via counting lesions using computer vision. In the third row, we propose to explore the label distribution for grading as introduced in section 3.1 via the correlation of these two tasks. The standard LDL assigns all the instances with label distributions of the same shape, while our proposed label distribution for the grading task is dynamically generated and achieves the grading accuracy of 82.05%. Then we combine the tasks of acne grading and lesion counting in the sixth row, which improves the performance on all the metrics. This indicates that the multi-task learning pattern based on the medical criterion can latently benefit the problem of acne severity grading. Furthermore, in the seventh line, the introduction of $L_{cnt2cls}$ loss and the average of the grading p_i^y and counting $\hat{p}_i^y$ results bring performance improvement to the model, because these two processes benefit the consistency between the counting and classification tasks and make the training and testing procedures of our method more stable.

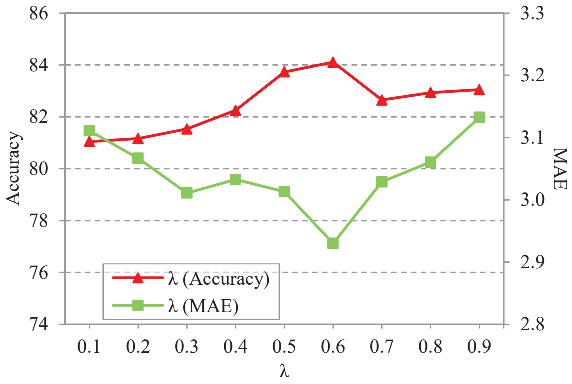

Figure 4. Model performance of our proposed method with different λ parameter.

Table 2. Ablation experiments demonstrating the effectiveness of different modules on the *ANCE04* dataset. "YI" indicates the Youden Index metric. "G" and "C" denote the grading and counting tasks, respectively. The values following '±' are standard deviations.

Method	MAE ↓	Precision ↑	YI ↑	Accuracy ↑
SLL(G)	-	75.81±2.56	67.21±4.11	78.42±2.11
LDL(G)	-	78.51±0.03	68.81±0.05	79.31±0.02
Ours(G)	-	80.56±0.02	71.67±0.05	82.05±0.02
SLL(C)	4.16±0.11	76.04±1.57	66.23±3.42	75.69±1.26
LDL(C)	3.24±0.14	80.39±0.02	70.65±0.01	81.09±0.01
Ours(G+C)	3.01±0.17	83.12±0.03	72.88±0.04	82.53±0.01
Ours	**2.93**±0.18	**84.37**±0.02	**75.32**±0.02	**84.11**±0.01

4.5. Comparison with Classification Methods

As shown in Table 3, we compare our method with three types of methods, including LDL, hand-crafted feature (HF), and deep feature (DF) based methods. The LDL methods contain PT-Bayes, PT-SVM, AA-kNN, AA-BP, SA-IIS, SA-BFGS, and SA-CPNN [22]. We extract feature representations from the last fully connected layer of the ResNet-50 [25]. Gao *et al.* [17] also propose a CNN-based deep LDL (DLDL) which is consistent with the LDL method in Table 2. Hand-crafted feature based methods consist of SIFT [30], HOG [7], GABOR [31], and color histogram (CH) [42]) representations. Extracted features are sent into a Support Vector Machine (SVM) classifier. Deep feature based methods consist of VGGNet-16 [39], Inception-v3 [43], and ResNet-50 [25].

We compare our method with several LDL methods as shown in Table 3. SA-BFGS [22] performs better than other classifiers, although the grading accuracy of 76.16 is lower than basic deep ResNet-50 [25] model. The DLDL performs the best because the DLDL trains an end-to-end CNN model. However, the standard label distribution, *i.e.*, assigning all the instances with the label distribution of the same shape, is more suitable for totally ordered labels such as the age. The labels of acne severity are ordered, yet it is too raw to represent the large intra-class variance. Our method explores the label distribution from the lesion counting task with continuous features for the acne severity grading and achieves the best performance.

Furthermore, compared with deep feature based methods, hand-crafted features perform poorly in all evaluation metrics. Different stages of acne have distinct variances between lesions in texture, color, and border. For example, acne lesions should have deeper color (*e.g.*, crimson or black) in the late stage compared with the early stage, while these low-level features are not enough to discriminate different acne severity. Since every severity level of acne will experience the same procedure from the early stage to recovery. Even for more severe acne, the lesions may present richer and clearer color *etc.* to a certain degree. The poor performance of YI metric also indicates that the diagnostic results from hand-crafted feature based methods have very low reference value.

In contrast, deep features represent the acne via high-level semantic information and perform better. ResNet-50 [25] achieves the best performance in basic CNN models (*i.e.*, the accuracies of 3.2% and 2.0% over the other two

Table 3. Comparison with the label distribution learning methods (PT-Bayes, PT-SVM, AA-kNN, AA-BP, SA-IIS, SA-BFGS, SA-CPNN, DLDL), hand-crafted feature based classification methods (SIFT, HOG, GABOR, CH), and deep methods (VGGNet, Inception, ResNet). The values following '±' are standard deviations.

Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	SA-CPNN	DLDL [17]
Precision	45.31±0.09	44.60±0.07	67.61±0.13	65.36±0.10	60.45±0.04	73.85±0.03	47.60±0.17	78.51±0.03
Specificity	79.39±0.03	83.04±0.03	87.73±0.07	87.37±0.02	85.93±0.01	91.01±0.01	80.40±0.03	92.24±0.01
Sensitivity	45.06±0.12	46.05±0.05	67.33±0.15	58.65±0.10	60.17±0.05	72.03±0.03	47.15±0.08	78.57±0.05
Youden Index	24.44±0.15	29.10±0.08	55.05±0.22	46.02±0.11	46.10±0.06	63.03±0.04	27.55±0.10	68.81±0.05
Accuracy	45.38±0.07	48.15±0.11	68.15±0.17	66.44±0.04	63.22±0.02	76.16±0.03	46.92±0.08	79.31±0.02
Criterion	SIFT [30]	HOG [7]	GABOR [31]	CH [42]	VGGNet [39]	Inception [43]	ResNet [25]	**Ours**
Precision	42.59±2.14	39.10±5.30	45.35±5.58	43.40±4.20	72.65±3.42	74.26±3.26	75.81±2.56	**84.37**±0.02
Specificity	78.44±1.10	77.91±1.53	79.89±1.58	78.70±1.06	90.60±0.71	90.95±0.68	91.85±0.77	**93.80**±0.00
Sensitivity	39.09±4.47	38.10±5.33	41.78±5.47	41.27±2.01	72.71±2.60	72.77±2.61	75.36±3.39	**81.52**±0.02
Youden Index	17.53±5.38	16.01±6.80	21.67±7.02	19.97±2.91	63.31±3.19	63.72±2.92	67.21±4.11	**75.32**±0.02
Accuracy	45.89±2.16	41.30±6.02	48.22±4.20	47.47±2.39	75.17±1.97	76.44±1.77	78.42±2.11	**84.11**±0.01

Table 4. Comparison with the state-of-the-art counting methods on the *ACNE04* datasets. The values following '±' are standard deviations.

Methods	MAE ↓	MSE ↓	Precision ↑	Specificity ↑	Sensitivity ↑	Youden Index ↑	Accuracy ↑
F-RCNN [35]	6.70±0.28	11.51±0.37	56.91±9.15	90.32±0.86	61.01±3.90	51.34±4.66	73.97±1.88
RefineDet [55]	5.82±0.53	10.14±0.49	72.20±1.70	89.53±0.60	66.03±5.10	55.56±5.69	72.09±1.46
YOLOv3 [34]	6.69±0.28	11.35±0.13	67.01±0.09	85.96±0.71	51.68±4.58	37.63±5.29	63.70±1.37
MCNN [56]	5.28±0.20	7.76±0.29	63.97±3.89	82.84±1.40	46.22±3.34	29.07±4.62	58.01±3.26
Ours	**2.93**±0.18	**5.42**±0.66	**84.37**±0.02	**93.80**±0.00	**81.52**±0.02	**75.32**±0.02	**84.11**±0.01

models respectively). Besides, it obtains at least 30% accuracy boost over hand-crafted feature based methods. Our method outperforms ResNet-50 in accuracy by a significant 5.69% and YI by 8.11%. This demonstrates the advantages of label distribution and multi-task learning strategies oriented by the professional medical criterion. The very low standard deviations further indicate the stability of our proposed method. Our method can benefit mining discriminative feature representation for acne images.

4.6. Comparison with Counting Methods

We compare with detection and regression based counting methods respectively. Table 4 illustrates the acne severity grading results, which are converted by counted lesion numbers based on the medical criterion [24], and lesion counting results respectively. Object detection methods such as Faster R-CNN [35] *etc*. generally outperform regression methods in the grading task due to the sparsity of acne lesions. However, detection based methods are unstable when the object size is small. For example, Faster R-CNN [35] achieves the best accuracy of 73.97% among the compared counting methods. But the poor performance such as precision and sensitivity indicates detection based methods are hard to achieve balanced results in each category. Specifically, as shown in Fig. 5, Faster RCNN (denoted as F-RCNN) achieves the accuracy of 12.44% on the acne severity of 'severe', while only achieving the accuracy of 4.62% on the acne severity of 'very severe'. The regression-based MCNN [56] achieves the best counting results especially in the MSE metric compared with detection based methods. However, the poor grading performance shows this method ignores the different weights between acne severity levels when counting lesions, *i.e*., the interval corresponding to different severity levels of acne is different. And our method can not only achieve excellent counting results but also balance the counting loss among different acne severity via label distributions. As shown in Table 4, our method outperforms compared methods both on classification and counting tasks. In addition, as illustrated in Fig. 5(b) and Fig. 5(c), our method achieves lower counting errors both on average and for each severity level of acne. Especially the significant improvement on the severity of 'very severe' demonstrates the stability of our method.

4.7. Comparison with Dermatologists

To validate the practical application value of our diagnostic system, we compare our method with 2 professional dermatologists and 2 general doctors. After getting familiarized with the Hayashi criterion [24], each doctor tests on 700 acne images which are randomly sampled from the

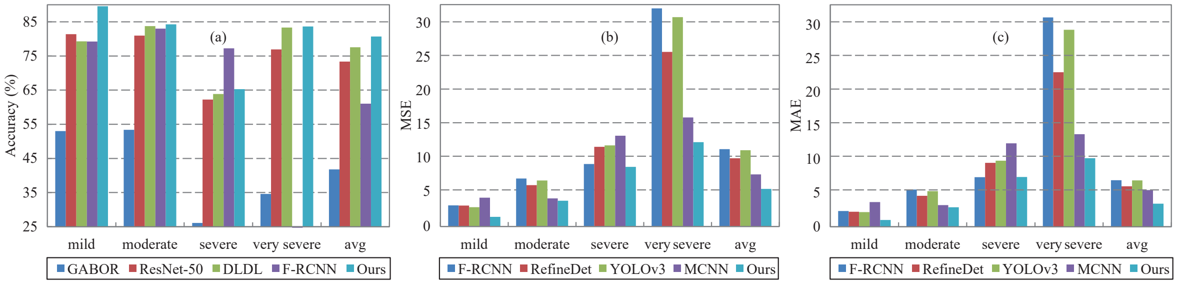

Figure 5. Classification (a) and counting (b, c) results of different methods on four severity levels of acne and average performance.

Table 5. Comparison of the doctors on the *ACNE04* datasets. 'Derm' denotes professional dermatologists who are knowledgeable in the dermatology field. 'GD' denotes general doctors who are not specialized in dermatology. The values following '±' are standard deviations.

Criterion	GD 1	GD 2	Derm 1	Derm 2	**Ours**
Precision	62.87	62.07	77.33	82.95	**84.37**±0.02
Specificity	84.11	86.98	90.66	92.16	**93.80**±0.00
Sensitivity	55.27	68.33	72.56	78.27	**81.52**±0.02
Youden Index	39.38	55.31	63.22	70.43	**75.32**±0.02
Accuracy	58.43	63.14	75.29	79.43	**84.11**±0.01

ACNE04 dataset. We report the grading results of each doctor in Table 5. We can observe that expert knowledge is of great importance in the acne grading procedure. General doctors show mediocre performance on each evaluation metric. The poor results of Youden Index value and accuracy indicate that they have a weak ability to clearly distinguish different severity levels of acne without expert knowledge. The dermatologists achieve better performance on all metrics. The gap in their results is due to the difference of acne diagnosis experience and personal subjectivity.

Our method achieves the dermatologist level performance and even exceeds the two dermatologists to a certain degree. This demonstrates that our method can provide valuable diagnosis evidence for doctors or patients. While the grading of acne severity is still a challenging task, as the examples illustrated in Fig. 6, the counting result can be regarded as the confident evidence for final acne grading. When the results of the classification and the number are inconsistent, they can also refer to each other such as Fig. 6(b). However, when both the results of classification and counting are wrong such as Fig. 6(c), we may need to introduce more prior expert knowledge or medical criteria to the diagnostic system.

5. Conclusion

In this work, we present a unified framework which can simultaneously learn to grade global acne severity and count local lesions. Oriented by the professional medical criterion, our method has two branches addressing grading and counting tasks. Our method learns the continuous feature representation of the acne image from the lesion counting task. Then it learns the severity label distribution to efficiently grade the acne image. To verify the effectiveness of the proposed method, we collect a dataset named *ACNE04*. We invite several experts to manually annotate the lesion bounding boxes and severity grade. Results indicate that our method can achieve dermatologist level performance and provide accurate diagnostic reference.

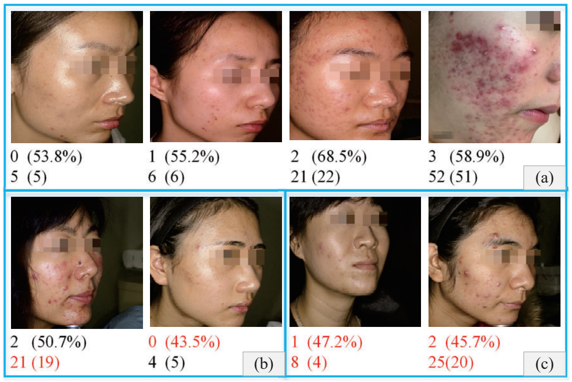

Figure 6. Examples of the counting and classification results. The numbers under the images denote the ground-truths of the acne severity and lesion number. The numbers in parentheses represent classification probability corresponding to true severity and estimated lesion number, respectively. The red font denotes the wrong prediction.

Acknowledgment

This work was supported by the NSFC (No. 61876094, U1933114), Natural Science Foundation of Tianjin, China (No. 18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities.

References

[1] Fazly Salleh Abas, Benjamin Kaffenberger, Joseph Bikowski, and Metin N Gurcan. Acne image analysis: Lesion localization and classification. In *SPIE*, 2016.

[2] Nasim Alamdari, Kouhyar Tavakolian, Minhal Alhashim, and Reza Fazel-Rezai. Detection and classification of acne lesions in acne patients: A mobile application. In *EIT*, 2016.

[3] Lokesh Boominathan, Srinivas S.S. Kruthiventi, and R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM MM*, 2016.

[4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018.

[5] Antoni B. Chan and Nuno Vasconcelos. Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012.

[6] Thanapha Chantharaphaichi, Bunyarit Uyyanonvara, Chanjira Sinthanayothin, and Akinori Nishihara. Automatic acne detection for medical treatment. In *ICTES*, 2015.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *CVPR*, 2018.

[9] Brigitte Dreno and Florence Poli. Epidemiology of acne. *Dermatology*, 206(1):7–10, 2003.

[10] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[11] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

[14] Deng-Ping Fan, Zheng Lin, Jia-Xing Zhao, Yun Liu, Zhao Zhang, Qibin Hou, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *arXiv preprint arXiv:1907.06781*, 2019.

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019.

[16] Maroni Gabriele, Ermidoro Michele, and Previdi Fabio. Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity. In *SSCI*, 2017.

[17] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[18] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, 2018.

[19] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE TPAMI*, 2019.

[20] Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *PR*, 2014.

[21] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, 2014.

[22] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[23] V. Goulden, G.I. Stables, and W.J. Cunliffe. Prevalence of facial acne in adults. *Journal of the American Academy of Dermatology*, 41(4):577–580, 1999.

[24] Nobukazu Hayashi, Hirohiko Akamatsu, Makoto Kawashima, and Acne Study Group. Establishment of grading criteria for acne severity. *The Journal of Dermatology*, 35(5):255–260, 2008.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[26] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image Processing*, 26(8):3846–3858, 2017.

[27] Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *AAAI*, 2017.

[28] Daniel P. Krowchuk. Managing acne in adolescents. *Pediatric Clinics of North America*, 47(4):841–857, 2000.

[29] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, 2018.

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[31] Rajiv Mehrotra, Kameswara Rao Namuduri, and Nagarajan Ranganathan. Gabor filter-based edge detection. *Pattern Recognition*, 25(12):1479–1494, 1992.

[32] Daniel Onoro-Rubio and Roberto J. López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, 2016.

[33] P.E. Pochi. The pathogenesis and treatment of acne. *Annual Review of Medicine*, 41(1):187–198, 1990.

[34] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[37] Payam Sabzmeydani and Greg Mori. Detecting pedestrians by learning shapelet features. In *CVPR*, 2007.

[38] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *AAAI*, 2019.

[41] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *ECCV*, 2016.

[42] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[44] Jasper R.R. Uijlings, Koen E.A. Van De Sande, Theo Gevers, and Arnold W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[45] Jing Wang and Xin Geng. Theoretical analysis of label distribution learning. In *AAAI*, 2019.

[46] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

[47] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 2018.

[48] Hywel C. Williams, Robert P. Dellavalle, and Sarah Garner. Acne vulgaris. *The Lancet*, 379(9813):361–372, 2012.

[49] Changdong Xu and Xin Geng. Hierarchical classification based on label distribution learning. In *AAAI*, 2019.

[50] Jufeng Yang, Liyi Chen, Le Zhang, Xiaoxiao Sun, Dongyu She, Shao-Ping Lu, and Ming-Ming Cheng. Historical context-based style classification of painting images via label distribution learning. In *ACM MM*, 2018.

[51] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017.

[52] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, 2017.

[53] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *CVPR*, 2018.

[54] Xu Yang, Xin Geng, and Deyu Zhou. Sparsity conditional energy label distribution learning for age estimation. In *IJCAI*, pages 2259–2265, 2016.

[55] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.

[56] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.

[57] Zhaoxiang Zhang, Mo Wang, and Xin Geng. Crowd counting in public video surveillance by label distribution learning. *Neurocomputing*, 166:151–163, 2015.

[58] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*, 2019.

[59] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.