16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Residual Squeeze-and-Excitation Network for Battery Cell Surface Inspection

Ziyang Song[1], Zejian Yuan[1], Tie Liu[2]
[1] Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University, Xi'an 710049, China
songzy305@yahoo.com, yuan.ze.jian@xjtu.edu.cn
[2] Information Engineering College
Capital Normal University, Beijing 100048, China
liutiel@163.com

Abstract

Anomaly detection remains a challenge in industrial inspection, due to difficulty in designing suitable features for classical methods and in collecting sufficient samples for deep learning based methods. We propose a novel Residual Squeeze-and-Excitation network to discover anomalies and inspect the quality of adhesives on battery cell surfaces. Owing to a compact architecture design and the utilization of an attention mechanism based module, our network generalizes well with a small amount of samples. A proper training setup further ensures our network a satisfying performance on a dataset constructed by ourselves. The network manages to accurately and robustly judge the existence of anomalies and adhesives and provide visual localization of them in an image of the battery cell surface.

1 Introduction

Automatic industrial inspection is one of the most important applications of computer vision technologies. Most of these inspection systems are centered on detection of surface anomalies in the industrial production environment. The appearance of anomalies, such as scratches, cavities and tarnishes, varies hugely. A qualified inspection system should be able to judge whether anomalies exist given an image of the object surface, and would be much appreciated if it could also provide the localization information of these anomalies.

Classical methods for anomaly detection often follow the same process, i.e., hand-crafted feature extractors followed by a trained classifier [1]. Hand-engineering feature descriptors are crucial for these methods. Deep learning based methods, which were early introduced for surface anomaly detection by [2] and [3], have shown significant advantages comparing to the classical methods, especially in tasks where feature descriptors are difficult to design. The training of a deep CNN often requires a large amount of training data which are usually not available since positive data samples, i.e., the images of object surfaces with anomalies are rarely seen and costly to collect compared to the normal samples. Such a problem commonly resides in scenarios where anomaly detection is applied. This situation motivates us to turn to a more reasonable design of our network architectures and training strategies.

In this paper, we focus on the detection of anomalies on battery cell surfaces. Those anomalies can cause the leak of chemical fluid inside battery cells and result in terrible accidents, thus makes the discovery of them a crucial and necessary task. We also perform an additional task of inspecting the quality of adhesives, i.e., glue coated on battery cell surfaces for subsequent sealing, simultaneously. Figure 1 helps illustrate our problem, which can be regarded as two independent classification tasks. Two scores between [0, 1] need to be predicted, with one indicating the likelihood of the existence of surface anomalies and the other suggesting that of adhesives. Although image classification is the most basic application of deep learning methods in computer vision, this task confronts a challenge which makes it different from simple image classification tasks: Both anomalies and adhesives appear to be mini-scale, accounting for few pixels even in images collected by an extremely high-resolution camera. Without the explicit guidance of localization of anomalies and adhesives, it is hardly possible for an image classification model to spontaneously concentrate on these sporadically distributed tiny structures, resulting in a poor training effect.

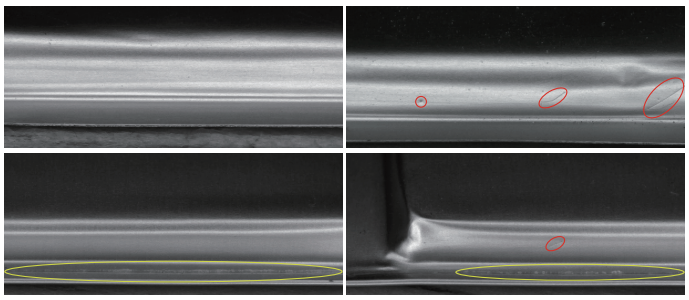

Figure 1. Examples of the battery cell surface (top left), with anomalies (top right), with adhesives (down left) and with both of them (down right).

Enlightened by [4], we split the model into a segmentation and a classification stage. Given annotated masks along with $\{0, 1\}$ labels, the model will be trained to segment the localization of anomalies and adhesives from image background and then use segmentation results to assist classification task. A novel Residual Squeeze-and-Excitation network is proposed, with a compact architecture design to learn from a handful of training samples and an attention mechanism based module to perform detection more efficiently and robustly. An optimum selection of loss

function in training also helps ensure us a satisfying performance of our approach on a dataset we construct by ourselves.

2 Residual Squeeze-and-Excitation Network

During the design of the architecture, several conditions below are mostly considered:

(1) surface anomalies occupy small local regions in images, thus require the network to keep detailed information which discriminates anomalies from surrounding background.

(2) Although adhesives appear as slim bars in images, they possess a unitary structure, which makes it feasible to segment out each fragment of a bar from surrounding regions without the need to capture a complex integral structure.

(3) Images with anomalies are very costly to collect in the real industrial production environment. The network should be able to fit well with a minimal amount of training samples.

Generally speaking, in the premise of capturing sufficient local receptive fields we should reduce the depth and complexity of the network as much as possible, due to the scarcity of training samples and the need for detailed information.

In recent years, the attention mechanism has been applied in many deep learning based methods, due to its powerful function to re-allocate the computation resource self-adaptively to most informative components in processing each sample [6][11][12]. The attention mechanism is especially useful for tasks with limited training data, including industrial inspection and anomaly detection tasks [10]. In order to utilize the computation resource and representation capacity of our compact network more efficiently, we propose an attention-like module to enhance the network, named "Residual Squeeze-and-Excitation" Module ("ResSE" for short in the following context).

Residual Squeeze-and-Excitation Module: The design of ResSE module is inspired by the Residual Attention [5] and Squeeze-and-Excitation [7] concepts. As shown in figure 2(a), the input to a ResSE module is a 3-D tensor $F^{in} \in R^{H \times W \times C}$, where H, W, and C representing the height, width, and channel dimension of a tensor respectively. We firstly perform a spatially global average pooling operation, namely *squeeze*, to aggregate spatially distributed information in each channel as $S \in R^C$:

$$S_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} F^{in}_{h,w,c} \tag{1}$$

The subsequent *excitation* operation aims to capture a channel-wise dependency relationship by modeling a nonlinear function of the spatially aggregated information attained by the previous step, formulated as:

$$E = Sigmoid(k_2 \times ReLU(k_1 \times S)) \tag{2}$$

where $k_1 \in R^{\frac{C}{r} \times C}$ and $k_2 \in R^{C \times \frac{C}{r}}$ refer to parameter matrix of two fully-connected layers. The bottleneck

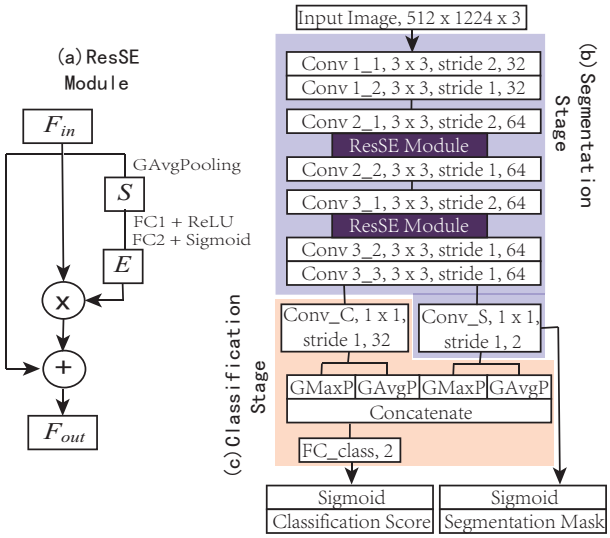

Figure 2. The proposed (a) ResSE module can be flexibly incorporated into our network, which consists of a (b) segmentation stage and a (c) classification stage. *Conv_S* in (b), and *Conv_C* and *FC_class* in (c) denote the segmentation layer, compression layer, and final classification layer respectively.

reduction ratio r (8 in our implementation) serves to control the number of additional parameters brought by this module. Finally, we use the *excitation* result $E \in R^C$ to modify the input tensor channel-wise and get the output of our ResSE module as:

$$F^{out}_{h,w,c} = (1 + E_c) \times F^{in}_{h,w,c} \tag{3}$$

conforming to the residual learning criteria that given the shortcut connection the performance should be no worse than its counterpart without residuals, which was firstly proposed in ResNet [8] and incorporated into attention mechanism in Residual Attention [5].

Segmentation Stage: We build a very shallow fully convolutional network architecture for the segmentation stage, as shown in figure 2(b). This stage is composed of three blocks, with the feature map spatially downsampled by a factor of two at the head of each block. The segmentation mask is predicted on the top feature map with a $1/8\times$ size of the input image. Given the common goal to capture tiny local structures, the anomaly and the adhesive localization tasks share the extracted feature maps. As a channel attention mechanism, the ResSE module can be incorporated into the segmentation stage to modify feature maps from any layer. Considering the tradeoff between improved accuracy and model complexity we design the architecture as shown in figure 2(b). Networks with various depths and with ResSE modules placed at other alternative positions in the segmentation stage are also evaluated. Experimental results will be discussed in section 4.3.

Classification Stage: In the classification stage we follow the architecture proposed by [4], which combines the top feature map and the segmentation result to predict a final classification score, as shown in figure 2(c). The compression layer serves to adjust the features extracted by convolutional layers above,

meanwhile reduce the model size to avoid overfitting with limited training samples. We jointly adopt maximum and average global pooling to extract information needed for classification from segmentation results and compressed features, since such a combination can robustify the classification score.

Note that although attention mechanisms can be integrated into a deep CNN in various ways, We merely import channel-wise attention into the segmentation stage. Our segmentation task is achieved by a fully convolutional neural network, thus makes spatial attention mechanism meaningless. No attention mechanism is adopted in the classification stage, because incorporating the segmentation result has provided explicit spatial attention. Besides this, the classification accuracy mostly relies on the quality of the segmentation task, which can be improved by bringing in our ResSE module in the segmentation stage. Furthermore, the classification stage itself contains rather few parameters, which makes the implementation of any attention mechanism to be costly and inefficient.

3 Training

The network training is split into two stages. In the first stage, we ignore the gradients produced by classification error and train the segmentation stage for 200 epochs. In the second stage, we freeze the parameters in the segmentation stage and run training on classification task for 80 epochs. Such a stage-wise training strategy provides the classification stage with sound segmentation results and meaningful feature representations, leading to a more stable and faster convergence. Note that during the training of the segmentation stage, unlike in [4] where the pixel-wise mean squared error is measured between the segmentation results through a linear activation and groundtruth masks, we append a *sigmoid* activation after the segmentation layer and minimize the pixel-wise binary cross entropy (CE) loss function, i.e.,

$$L_{segm} = \frac{1}{BHW \times 2} \sum_{b=1}^{B} \sum_{p=1}^{H \times W} \sum_{c=1,2} CE(\widehat{y}_{b,p,c}, y_{b,p,c}) \tag{4}$$

We make this modification for the following two reasons: (i) We prefer the segmentation stage learn a binary distribution rather than regress a continuously distributed score for each pixel; (ii) Without a sigmoid activation the output of the segmentation layer fed into the classification stage will be restricted between $(0, 1)$ or $(-1, 1)$ after the first stage training, thus weaken its contributions to the classification stage. Experiments in section 4.3 will prove the superiority of our choice.

In the training of the classification stage, the traditional binary cross entropy (CE) loss function is minimized, i.e.,

$$L_{class} = \frac{1}{B \times 2} \sum_{b=1}^{B} \sum_{c=1,2} CE(\widehat{y}_{b,c}, y_{b,c}) \tag{5}$$

In both stages loss functions are minimized on a batch size of 16 using the AdamOptimizer [9] with the default parameter settings suggested in the paper.

4 Experiments

4.1 The Dataset

We construct a dataset by ourselves to evaluate the network. The images of industrial battery surface are collected by a high-resolution camera, under which a pixel approximately corresponds to a square with sides of length 7 μm. As shown in figure 1, adhesives on the surface present a rather unitary pattern, while anomalies appear in various forms like scratches, cavities, and traces left by corrosions.

The entire dataset consists of 591 train examples and 323 test examples with 512×1224 size and RGB channels. For each sample, the localization of these anomalies and adhesives are roughly annotated with points or broken lines and two groundtruth masks are later generated according to these annotations. A value pair $y \in \{0, 1\}^2$ denotes the existence of anomalies and adhesives in each sample.

Data Augmentation: In real industrial production scenarios, the distance from the camera to the battery surfaces can be versatile, thus results in multi-scale views of battery surfaces. Given samples which are collected by the camera from a fixed distance away, we propose a data augmentation method to simulate real scenarios and generate samples in various scales during training. For each sample, a crop ratio δ no less than the pre-defined threshold Δ (0.75 in our implementation) is randomly determined. Then a patch with the same shape and $\delta\times$ size as the original image is randomly cropped from the original image. The cropped patch is resized to a uniform input size and fed into our network.

4.2 Setup

All of our experiments are ran with the same configurations. Given an RGB input image of size 512×1224, our network outputs two segmentation masks of size 64×153 and two classification scores for surface anomalies and adhesives. The whole training process costs 4 hours on a NVIDIA GTX 980 GPU with 4GB memory. In the prediction, the network can reach a speed of 25 fps with the same equipment.

4.3 Results

We evaluate the performance of our network in terms of the classification robustness, namely the performance with various classification threshold, for both positive and negative samples for anomalies and adhesives. The Average Precision (AP) metrics are measured for the trained network. We also evaluate the performance of our network in terms of the true positive rate and true negative rate, i.e., the proportion occupied by ones which are correctly identified in positive and that in negative test samples, under a specified classification threshold which we set to be 0.7 as most of classification tasks do in real scenarios.

Network Depth Selection in the Segmentation Stage: We firstly remove the ResSE modules and test the performance of baseline networks with various depths. The segmentation stages of these baseline networks all conform to the three-block design rule mentioned in section 2, while differ in the number of layers

Table 1. Architectures of various baseline networks with a uniform three-block design.

network	number of layers within each block
baseline v1	1, 1, 2
baseline v2	2, 2, 3
baseline v3	3, 3, 4
baseline v4	4, 4, 5

within each block as illustrated in table 1. As can be seen in table 2, from *baseline v1* to *v2* we obtain more powerful features for segmentation and final classification by increasing the network depth. However, upgrading to *baseline v3* only provides slight improvements and *baseline v4* even brings a drop in performance due to overfitting. Experimental results prove that the depth and complexity of our baseline network design, namely *baseline v2*, fit well with the limited amount of training data and meanwhile can save the computational resource as much as possible. ResSE modules should be incorporated into *baseline v2* network (*baseline* for short in the following context) for better performance and efficiency.

Table 2. Performance of various baseline networks.

Performance on anomaly detection

network	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
baseline v1	91.7	18.6	83.0	68.1
baseline v2	93.1	40.4	85.9	79.5
baseline v3	**93.6**	**41.2**	**87.2**	79.5
baseline v4	93.3	36.5	86.8	**81.8**

Performance on adhesive detection

network	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
baseline v1	99.3	90.8	92.7	100.0
baseline v2	99.3	91.8	92.7	100.0
baseline v3	**99.4**	**92.5**	91.4	100.0
baseline v4	98.4	91.3	**94.7**	100.0

Benefits from the ResSE Module: We make a comparison among networks with various architectures regarding their performance. Table 3 illustrates their respective structure designs. As shown in table 4, all the networks with additional ResSE modules surpass the baseline network more or less, while the *ResSE v1* network performs best. Such a result accords with our theoretical expectation, since the channel attention mechanism has little effect while being placed at bottom layers due to the close inter-channel correlations at that stage (*ResSE v2*), and has no opportunity to be effectively utilized by subsequent layers if being placed at top layers (*ResSE v3*). Therefore, we choose this version for subsequent experiments and real applications.

Loss Function Selection in the Segmentation Stage: We have analyzed the loss function setup for the segmentation stage training in section 3. Here we test the effects of different loss functions on *ResSE v1* network and the results conform to our theoretical analysis, as shown in table 5.

Qualitative Analysis: We further perform a qualitative analysis of our network in terms of its segmentation results. Several test samples are depicted in figure 3 and 4, with annotated and predicted foreground masks pasted onto them. Adhesives are often easy to identify due to their unitary structure. Our network can robustly recognize and localize anomalies with various appearance, as shown in figure 3.

However, because of their diversity in appearance, our network fails to identify some of the anomalies which are rare in training samples like in figure 4(a) and 4(b). Besides this, some artifacts like wrinkles or tarnishes, which should not be classified into anomalies, are segmented out due to their distinction from surrounding areas like in figure 4(c) and 4(d). More complex and deeper architectures seem to be required for capturing such structures. Therefore, our future work should concentrate on collecting more data samples to increase our generality and optimizing the network architecture design concerning the tradeoff between mini-scale structure and large-scale complex structure identification.

Table 3. Structure designs of various architectures, including *baseline*, i.e., the network without ResSE module incorporated.

network	layers followed by ResSE modules
baseline	*without ResSE modules*
ResSE v1	Conv 2_1, Conv 3_1
ResSE v2	Conv 1_2
ResSE v3	Conv 3_2, Conv 3_3

Table 4. Performance of various architectures.

Performance on anomaly detection

network	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
baseline	93.1	40.4	85.9	79.5
ResSE v1	**94.7**	**41.4**	**89.8**	**89.8**
ResSE v2	93.4	41.2	87.7	78.4
ResSE v3	93.8	40.9	86.4	81.8

Performance on adhesive detection

network	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
baseline	99.3	91.8	92.7	100.0
ResSE v1	**99.7**	**98.8**	**96.1**	100.0
ResSE v2	99.5	**98.8**	93.4	100.0
ResSE v3	99.2	94.2	93.4	100.0

Table 5. Effects of various training strategies.

Performance on anomaly detection

activation + loss	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
linear + MSE	90.6	5.8	81.7	71.6
sigmoid + MSE	94.6	35.7	87.6	88.6
sigmoid + CE	**94.7**	**41.4**	**89.8**	**89.8**

Performance on adhesive detection

activation + loss	AP_{pos}	AP_{neg}	$TPR_{0.7}$	$TNR_{0.7}$
linear + MSE	99.6	87.1	92.1	100.0
sigmoid + MSE	99.6	98.2	94.7	100.0
sigmoid + CE	**99.7**	**98.8**	**96.1**	100.0

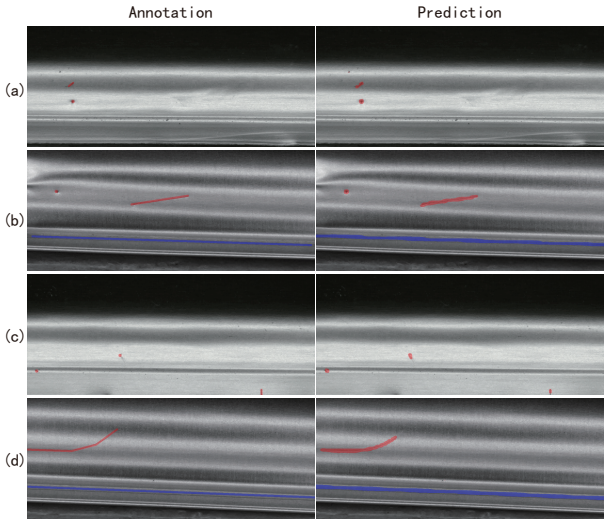

Figure 3. Examples of test samples that are correctly localized and classified by our network.

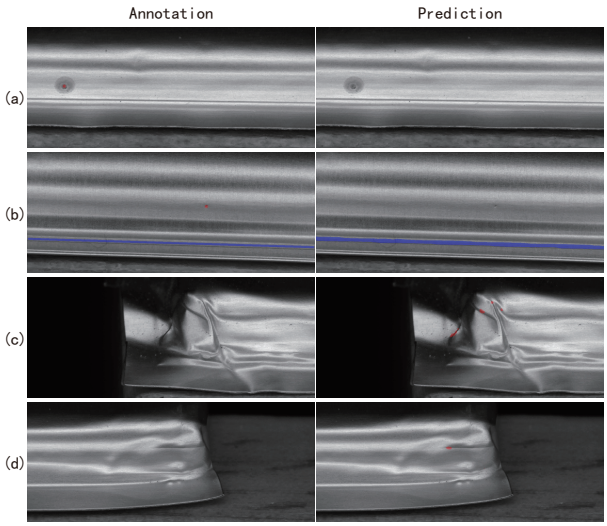

Figure 4. Examples of test samples that are wrongly localized and classified by our network.

5 Conclusion

In this work, we propose a novel ResSE network based on deep convolutional neural network for the inspection of anomalies and adhesives on battery cell surfaces. Concerning the characteristics of the data samples, we design a very compact network architecture and import ResSE modules to utilize its representation capacity and computation resource more efficiently. The network achieves satisfying performance on our dataset by accurately predicting the existence and localization of anomalies and adhesives in sample images. Due to its compactness and effectiveness, we believe that our work can be easily transferred to a similar domain with slight configuration and hyperparameter adjustments.

6 Acknowledgments

This work was supported by the National Key RD Program of China (No.2016YFB1001001), the National Natural Science Foundation of China (No.91648121, No.61573280, No.61603022).

References

[1] X. Xie.: "A Review of Recent Advances in Surface Defect Detection using Texture Analysis Techniques," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol.7, no.3, pp.1-22, 2008.

[2] J. Masci, U. Meier, D. Ciresan, et al.: "Steel Defect Classification with Max-pooling Convolutional Neural Networks," *The 2012 IEEE International Joint Conference on Neural Networks (IJCNN)*, pp.1-6, 2012.

[3] D. Weimer, B. Scholz-Reiter, and M. Shpitalni.: "Design of Deep Convolutional Neural Network Architectures for Automated Feature Extraction in Industrial Inspection," *CIRP Annals-Manufacturing Technology*, vol.65, no.1, pp.417-420, 2016.

[4] D. Racki, D. Tomazevic, and D. Skocaj.: "A Compact Convolutional Neural Network for Textured Surface Anomaly Detection," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[5] F. Wang, M. Jiang, C. Qian et al.: "Residual Attention Network for Image Classification," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] W. Li, X. Zhu, and S. Gong.: "Harmonious Attention Network for Person Re-Identification," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] J. Hu, L. Shen, and G. Sun.: "Squeeze-and-Excitation Networks," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] K. He, X. Zhang, S. Ren, et al.: "Deep Residual Learning for Image Recognition," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] D. Kingma, J. Ba.: "Adam: A Method for Stochastic Optimization," *Open Access Library (OALib)*, 2014.

[10] J. Zhu, Z. Yuan, and T. Liu: "Welding Joints Inspection via Residual Attention Network," *16th International Conference on Machine Vision Applications (MVA)*, 2019.

[11] K. Gregor, I. Danihelka, A. Graves, et al: "DRAW: A Recurrent Neural Network For Image Generation," *32nd International Conference on Machine Learning (ICML)*, 2015.

[12] J. Ba, V. Mnih, and K. Kavukcuoglu: "Multiple Object Recognition with Visual Attention," *The International Conference on Learning Representations (ICLR)*, 2015.