Projet de Recherche (PRe)

Spécialité : STIC
Année scolaire : 2019/2020

Tomography image quality assessment

Mention de confidentialité
Rapport non confidentiel

Auteur : Kai ZHANG
Promotion : 2021
Tuteur ENSTA Paris : Hammami Omar
Tuteur organisme d'accueil : Arnaud Demortière

Stage éffectué du 18 / 05 / 2020 au 21 / 08 / 2020

NOM de l'organisme d'accueil : Université de Picardie Jules Verne
Adresse : Chemin du Thil - CS 52501 - 80025 Amiens Cedex 1, France

CONFIDENTIALITY NOTICE

This report, written by ZHANG kai, student of ENSTA Paris, under the supervision of Arnaud Demortière, materials scientist of Laboratoire de Réactivité et Chimie des Solides (LRCS), CNRS UMR 7314, is non-confidential and may be published online.

Acknowledgements

First, I would like to thank Arnaud Demortière, the supervisor of my internship, for giving me the wonderful opportunity to conduct my internship within his research group. I have learned a lot from the routine working experience not only the research skills but also the communication skills. Besides, I want to give great thanks to Tuan-Tu Nguyen, who gave me lots of advice during my internship and offers the help for my stay in Amiens.

I also want to express my gratitude to Hammami Omar, my tutor of ENSTA Paris, who gave me the care on both living and study during the difficult period caused by COVID-19.

Finally, a sincere appreciation to my families for their support and care. Also, the thanks to the China Scholarship Council for the grant during my study in France.

Résumé

La perception des images joue un rôle fondamental dans les approches basées sur la tomographie pour la caractérisation de la microstructure et a un impact profond sur toutes les étapes ultérieures de traitement d'image (segmentation et analyse). Cependant, l'amélioration de la perception de l'image implique fréquemment la dépendance de l'observateur, qui se traduit par une dispersion d'utilisateur à utilisateur et des incertitudes dans les paramètres calculés. Ce travail présente une méthode quantitative objective, qui utilise des réseaux de neurones convolutifs, pour l'évaluation de la qualité de l'image tomographique. Par rapport à la plupart des méthodes basées sur les données existantes, notre méthode nécessite moins d'annotations et est plus appropriée pour les applications d'images tomographiques. Différentes mesures ont été utilisées pour évaluer la corrélation de nos scores prédits avec l'opinion humaine subjective ainsi que la précision de la segmentation. Les résultats de l'évaluation de ce travail démontrent que notre méthode peut être un outil direct qui guide le processus de valorisation et conduit à une segmentation fiable des résultats par rapport à l'opinion humaine subjective. En conséquence, le traitement d'image peut se transformer en un processus très robuste et indépendant de l'observateur.

Mots-clés:

Évaluation de la qualité d'image, segmentation sémantique, images tomographiques

Abstract

Images perception plays a fundamental role in the tomography-based approaches for microstructure characterization and has a profound impact on all subsequent image processing steps (segmentation and analysis). However, the enhancement of image perception frequently involves the observer-dependence, which translate into user-to-user dispersion and uncertainties in the calculated parameters. This work presents an objective quantitative method, which utilizes convolutional neural networks, for tomographic image quality assessment. Compared to most existing data-driven methods, our method requires less annotations and is more appropriate for tomographic images applications. Different metrics were employed to evaluate the correlation of our predicted scores with the subjective human opinion as well as the segmentation accuracy. The evaluation results from this work demonstrate that our method can be a direct tool that guides the enhancement process and conduct to a reliable segmentation results in respect to the subjective human opinion. As a result, the image processing can turn into a very robust, observer-independent process.

Keywords:

Image quality assessment, Semantic segmentation, Tomography images

Table of Contents

List of figures

List of tables

Chapter 1 - Introduction

I.1. Tomography image analysis pipeline

X-ray tomography has been considered as a powerful technique for studying lithium ion batteries (LIBs) since its nondestructive 3D imaging across multiple length scales, which provides quantitative or qualitative insight into battery operation and degradation. When the X-rays pass through the battery, they are partially attenuated. A scintillator converts the transmitted X-rays into visible light that can be imaged with a camera. An image taken at one fixed angle of the rotation stage yields a projection image that provides information about the absorption of the battery in that particular orientation. With numbers of projection images taken at different orientation, a tomographic reconstruction algorithm is applied to construct the 3D volume that provides the attenuation data for each point in the volume. Due to the different attenuation coefficients of each phase in the battery, the internal structure is visible [1].

However, as the shortcomings in the image acquisition process, the images suffer from kinds of distortions [2]. The most common impairments are: (a) the noise caused by the low count of incoming radiation at the detector, (b) image blur due to movement, hardware constraints, or suboptimal image reconstruction and (c) ring artifact affected by the defects of the camera or scintillator.

Generally, the pipeline of analyzing the microstructural properties through the tomography images consists of the image preprocessing, which aims at improving the image by removing distortions, and segmentation, which helps separate the regions to different phases. In image pre-processing procedure, several algorithms are utilized for image contrast improvement (histogram normalization, equalization, brightness adjustment etc.) and distortion suppression (median filters, ring artifact removal, unsharp mask etc.) [3]. The processed images are assessed by humans and the one with the best perceptual features is picked out for further segmentation. To create the segmentation mask, the most common method is the threshold-based algorithm (Otsu [4], Fuzzy-c means [5]), which separates the regions by comparing the pixel values with a threshold and assigns the class labels to corresponding regions. The quality of the segmentation is judged by visual inspection as well, due to the lack of ground truth masks. In total, the analysis procedures are very subjective because the quality and accuracy are assessed by visual inspection. What's worse, there is no ideal image for reference. It may lead to conflicts for the same image due to the different human observers with various background knowledge. Moreover, it is difficult to convince people of some specific method is better than the others without an objective metric.

Several experiments have been performed to investigate the impact of distortion on further analysis. Schluter et. al. [3] analyzed the segmentation

accuracy of images before and after preprocessing, which included denoise and ring artifact removal algorithms, and pointed out that the blur distortion resulted in the poor segmentation. Martin et al. [6] carried out the reconstruction experiments on computed tomography images with different levels of noise and the qualitative results revealed that the noise leaded to the decline in reconstruction quality.

In summary, an objective metric for assessing the image quality is urgently required not only for measuring the performance of different image processing methods but also for minimizing the damage of distortion on further analysis.

I.2. Image quality assessment

I.2.1 Description

Image quality assessment (IQA) aims to predict the perceptual quality of a distorted image. However, human vision system (HVS) needs a reference to quantify the discrepancy by comparing the distorted image either directly with the original undistorted image, or implicitly with a hallucinated scene in mind. It is time-consuming and labor-intensive to assess image quality from a crowd of people. Additionally, due to the different cultures and living backgrounds, people prefer taking different views to the same image. Especially for tomography images, the experts and the people without any prior knowledge would like to give totally different scores.

To avoid the distinction of results caused by various background knowledge and provide the professional estimation, some machine assisted IQA methods were proposed during the past few decades. They are generally divided into three categories: 1) full-reference image quality assessment (FR-IQA) which assesses the distorted image by comparing with the reference image and measuring the difference [7], [8]. 2) reduced-reference image quality assessment (RR-IQA) which measures image quality with part of the reference image [9], [10]. 3) no-reference image quality assessment (NR-IQA) which requires nearly no information on the reference images and estimates the image quality directly from distorted images [11], [12].

The following sub-sections describe the FR-IQA briefly. Due to the limited usage of RR-IQA, we would surpass the RR-IQA introduction and directly to the NR-IQA methods, which are more extensively studied.

I.2.2 FR-IQA

The conventional metrics used for FR-IQA are the peak signal-to-noise ratio (PSNR) and the mean square error (MSE) which compare the distorted images with the reference images directly on the intensity of the images without considering the HVS. By considering the luminance, contrast and structural information, SSIM[7] used average pooling to calculate a score from similarity map. Based on SSIM, MS-SSIM[13] compared the distorted image with the

reference image on multiple scales. F-SIM[14] leveraged phase congruency and gradient magnitude features to derive a quality score while GMSD[15] only considered the image gradient as the criterial features. Besides the gradient, MDSI[16] utilized chromaticity similarity and deviation pooling to imitate the HVS and achieved better results.

I.2.3 NR-IQA

As the lack of the reference images, the NR-IQA methods are more practical for in real-world applications. Generally, the NR-IQA methods can be divided into two categories: the distortion-specific method which is based on the prior knowledge of the distortion type; the non-distortion-specific method which evaluates the distorted images without any information about the distortion type. For example, focusing on evaluating the quality of images with noise and blur, Chen et al. [17] combined the statistics of the image gradient histogram and the image wavelet decomposition result to assess the blur images. Alexandre et al. [18] integrated many metrics into a paradigm for robust quality assessment. Niranjan [19] proposed a metric to estimate the cumulative probability of noticeable blur area and calculate the quality score. Qing [20] built a triangle model to extract the features of gradient profile sharpness and conducted a blur assessment metric based on the model.

Due to the limitation of prior knowledge about distortion type and the fact that distortions exist together, non-distortion-specific methods are preferred in real-world applications. These methods estimated the quality score through the assumption that the predicted distortion has the similar type to one of those in the training dataset. Correspondingly, these methods followed the two approaches [21]: (1) natural scene statistics (NSS) based method and (2) learning based method. Under the concept that natural images have strong statistical regularities among different visual contents, NSS method will capture the statistical properties and compare the distance between distorted images and good-quality images [22]. Due to the process of modeling the statistical features in a dataset, the single NSS method tends to be time-consuming. To reduce the computation burden, Michele et al. [23] proposed a blind IQA method based on the NSS model of discrete cosine transform coefficients rather than the original image, and the quality score was derived from the parameters of the model. In addition, the Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) [24] firstly identified the distribution type and then performed the assessment measure of perceptual quality based on the NSS wavelet coefficient model. Anish et al. [25] utilized the scene statistics of locally normalized luminance coefficients to obtain the losses of naturalness of the distorted images. This method had very low computational complexity and was practical for real time applications.

With the development of deep neural network (DNN) technology, the deep learning methods were exploited to learn the distortion types and estimate the image quality scores. Le et al. [26] firstly proposed a shallow convolutional neural network to estimate the image quality score. Ke [27] introduced a deep learning-based image quality index for blind image quality assessment, which was more efficient and robust. Instead of the multi-stage methods, Sebastian et al. [28] presented an end-to-end neural network to regress the quality score by joint

learning of local quality and local weights. Instead of considering the whole image in the network, Simone et al. [29] cropped the image patches, estimated the scores separately and fused them finally which was more suitable when there was not sufficient training data. However, the lack of training data was a crucial limitation to aforementioned method. To overcome the limitation of data, Xialei et al. [30] implemented data augmentation by generating artificial distorted images and then trained a Siamese network to regress the scores, which achieved outperformed results. Kwan-Yee et. al. [31] combined the generative neural network to generate the reference images and convolutional neural network to regress the quality score from the discrepancy.

I.3. Summary

Although many DNN methods have been provided for IQA and achieved excellent results, most of them require huge number of annotated labels, which are not suitable for tomography images. The already developed NR-IQA methods required less data than the FR-IQA method, but they still demanded for relatively large amount of data (hundreds of annotations). Besides, the existing open-source datasets [32] of battery electrodes tomographic images are not for the purpose of IQA task (i.e. without various of distortion-types and corresponding scores). Therefore, a light NR-IQA method which requires less annotated data and is robust to transfer among different tomography images is urgently demanded.

The main contributions of this work are summarized as follows:

- A NR-IQA method is proposed for TIQA which requires only dozens of annotated images and achieves outperformed results.
- A data generation method is developed by imitating the human observers to label the distorted images automatically for the purpose of addressing the insufficient data problem.
- A segmentation experiment is conducted to analyze to the correlation between our predicted quality scores and the segmentation performance.

The remainder of the paper is organized as follows: In chapter 2, we demonstrate the results of each module in our TIQA method and the relation between quality score and segmentation performance. In chapter 3, we summarize the results and emphasize the features of our method. In chapter 4, we introduce our approaches in details. In chapter 5, we make the conclusion.

Chapter 2 - Results

II.1 Data generation

As shown in Figure 1, the first step in our approach is to generate the data that is required for the following training process of the score prediction network.

Data generation includes: (i) distorted image generation, and (ii) label projection to infer the annotation of distorted image based on the difference of HVS features between original image and distorted image.

As illustrated in Figure 2, the original image is firstly resized and cropped into a fixed size, 224*224 pixels, in our experiments to assure the network can perceive the images in the same scale. To verify if the resize operation affects the image quality scores, we compare the annotations on the images before and after the resize operation. The comparison results (Supplementary Figure S1) confirmed that this operation has no effect on the annotated scores. Three types of distortion are applied, including ring artifact, blur, noise, to create distorted images. (Readers can refer to the Figure S2 for more examples on distorted images generated for this work).

Before introducing the results, we would like to indicate several metrics used in the evaluation procedure. For IQA results evaluation, some metrics, including Pearson's linear correlation coefficient (PLCC), Spearman's rank ordered correlation coefficient (SROCC) (More details are explained in section “Methods”) are applied to measure the correlation between predicted results and corresponding labels.

The evaluation of label scores and generated scores is shown in Figure 3 and Table 1, which indicate that regardless the type of distortions, the generated scores have the similar trend with the human labels extracted from our surveys. Especially for the images with noise or blur, from the SROCC and PLCC values, the generated scores have relatively high correlation with the human annotations. As for the ring artifact, the results demonstrate that the general FR-IQA metrics cannot well handle this type of distortion.

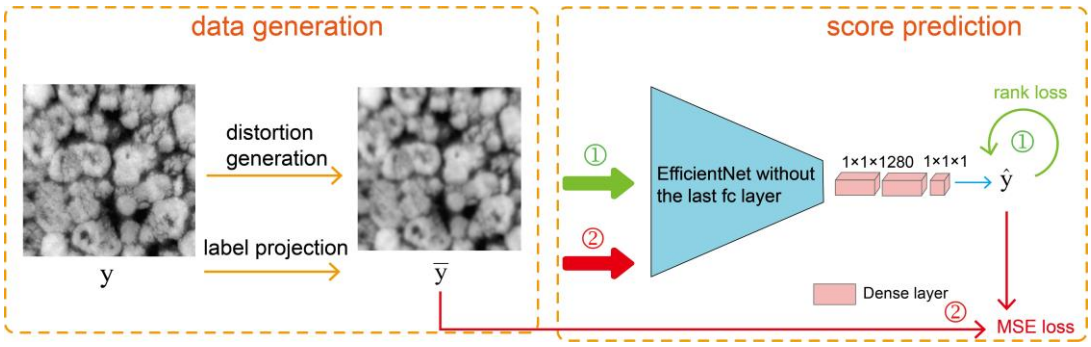


Figure 1 Pipeline of our TIQA method. It is composed of two modules: the data generation and score prediction. In score prediction, ① is the self-supervised learning for ranking the images, ② is the fine tune procedure for regressing the ranks to a score in fixed range.

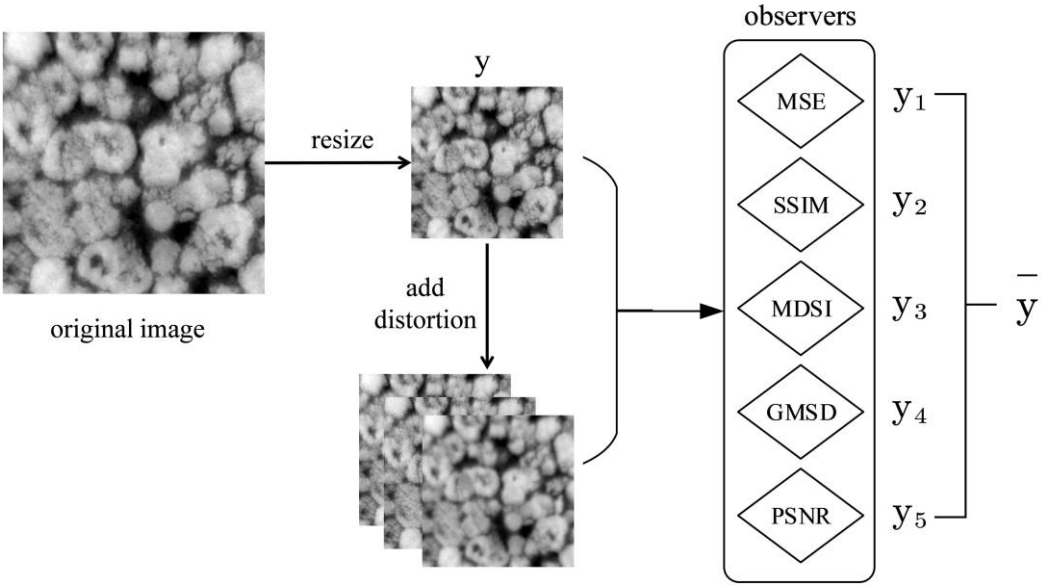


Figure 2 Detailed structure of data generation. The observers consist of five FR-IQA evaluators. y is the human annotation, y_i is the predicted score of the *i-th* observer, $\bar{y}$ is the average score of y_i.

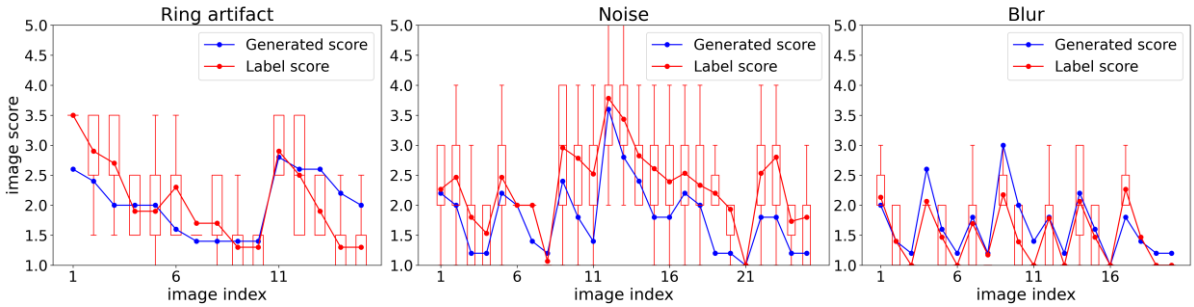


Figure 3 The qualitative results of label projection. The red box means the human labels with confidence. The red dot in each line is the average value of labels. In y axis, the image score ranging from 1 to 5 corresponds to the five quality levels: terrible, bad, average, good, excellent.

metric / distortion	Num of images	SROCC	PLCC
Ring artifact	15	0.655	0.641
Noise	25	0.813	0.858
Blur	20	0.852	0.846

Table 1 Quantitative evaluation results of label projection module.

II.2 Score prediction results

In the procedure of the image quality scores prediction, as shown in Figure 1, the network was firstly trained to learn to order the images according to the distortion level. Then based on the prior ranking knowledge, it was finetuned to regress the rank to a comprehensive quality score in a fixed range.

We take the EfficientNet[33] as the feature extractor instead of VGG[34] used in RankIQA[30] because it has less parameters, about 9 million, than VGG, about 138 million trainable parameters, which means easier to converge and less possibility for overfitting. With the trained model, we predicted the quality score on test set with 56 images and compared the results with human annotations.

As illustrated in Figure 4, the results on images with different types of distortion were evaluated separately. In total, these two lines follow the similar trend, which demonstrates our method works well in estimating the relative order of image pairs. Especially for blurred images, it performs excellently on imitating the HVS to predict the quality score.

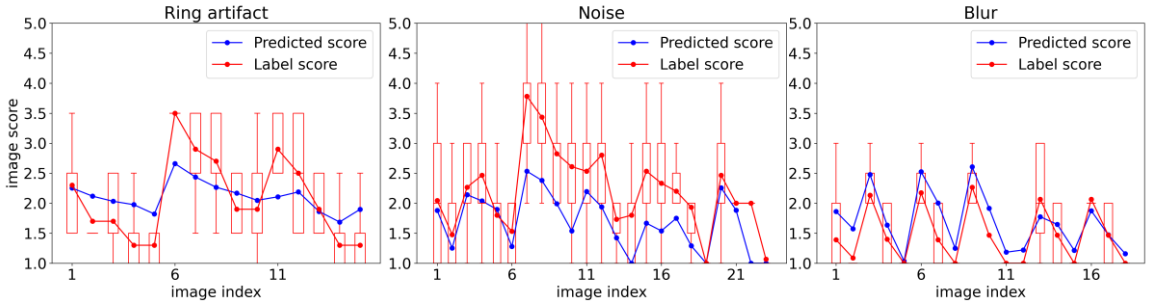


Figure 4 Evaluation results of score prediction module. The figures from left to right are the evaluation of images with ring artifact distortion, noise and blur, respectively. The red box represents the scores with confidence.

We compared our method with BRISQUE [25], RankIQA, and the quantitative evaluation results are presented in Table 1. Totally, from the column ALL, it shows that our method excels at assessing the quality of tomography images. Our method achieves the best accuracy among the three methods. In terms of the different types of distortion, our method outperforms BRISQUE for all the three distortions. When compared with RankIQA, our method achieves better

results for images with ring artifact and noise, and obtains nearly the same performance on the images with blur distortion.

Metric	Method	Ring Artifact	Noise	Blur	ALL
SROCC	BRISQUE[25]	0.785	0.737	0.861	0.794
	RankIQA[30]	0.761	0.769	**0.897**	0.809
	TIQA(ours)	**0.839**	**0.778**	0.895	**0.837**
PLCC	BRISQUE[25]	0.450	0.757	0.849	0.685
	RankIQA[30]	0.758	0.768	**0.879**	0.801
	TIQA(ours)	**0.871**	**0.799**	0.854	**0.841**

Table 2 Performance evaluation on our testing dataset. RankIQA uses the VGG as feature extractor while our method TIQA uses the EfficientNet. The rows marked light blue are our results. The bold values in each column are the best results. The column named ALL means the performance in all types of images.

II.3 Relation between IQA results and segmentation accuracy

To inspect the relation of the image quality and segmentation accuracy, we implemented a segmentation network to obtain the semantic segmentation on the test dataset, as illustrated in Figure. 5. Finally, the IQA results were compared with the segmentation performance to explore the influence of the distortions.

The F1 score and IoU score, which are calculated from confusion matrix[35] are taken to describe the segmentation accuracy. The qualitative results of TIQA method and segmentation are illustrated in Figure 6. We picked an original image and its corresponding images with different types of distortion as the data for both IQA and segmentation. From the results, we can see that the distortion affects the image quality and the segmentation performance. With the distortion, images obtain lower quality score and F1 score which means worse segmentation accuracy. The uncertainty map and IoU map clearly present the influence of distortion on final classification results. Compared these three types of distortion, the noise has severe effects on the results through causing error holes. Although it seems that the blur distortion has little uncertainty, it makes the boundaries vague and leads to more misclassification and reduction in HVS.

The evaluation results of the IQA and segmentation accuracy are shown in Figure 7 and Table 2. From the SROCC and PLCC, they prove that the quality scores predicted by our approach are related to the segmentation accuracy. The image quality score shares the similar trend with the segmentation accuracy, especially for the images with ring artifact and blur. From Figure 7, we can come to the same conclusion because the colorful lines (with distortion) are close to the black line (without distortion). Nevertheless, they may have different sensitivity for specific type of distortion. For example, the SROCC and PLCC score for images

with noise imply that human may react slightly different to the noise distortion. The point correlation line of the segmentation masks with noise do not converge at the reference line. The fluctuation indicates the severe impact of noise on segmentation results. Additionally, the third figure in Figure. 7 illustrates that with the increase of distortion level, the IQA score decreases quickly but the segmentation accuracy is stable, which means the network can tell the very little difference of pixel values in the image and classify the pixels to different categories on the basis of the distinction. But due to the limitation of HVS, people cannot distinguish the little variation of the pixel values.

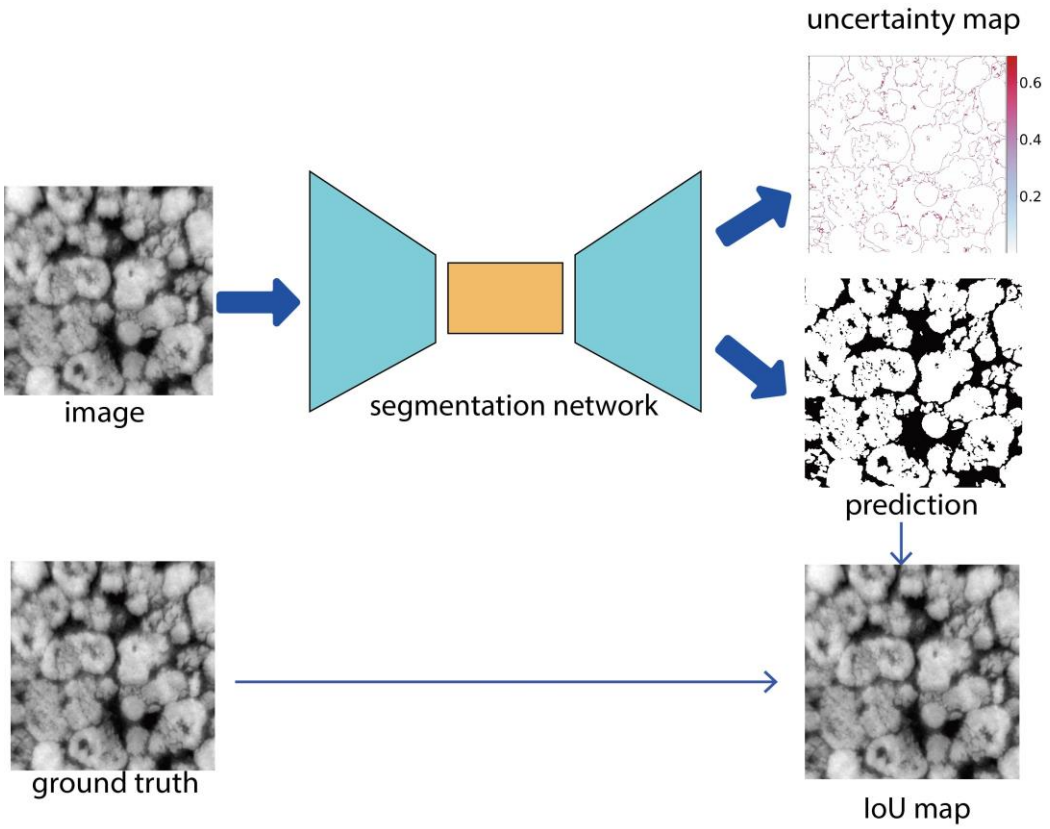


Figure 5 Pipeline of segmentation evaluation procedure.

In summary, through the image quality score produced by our method, especially for images with noise and ring artifact, we can infer the corresponding segmentation performance without implementation. It greatly reduces the time of choosing the appropriate algorithm of preprocessing the image for improving the quality to achieving better segmentation accuracy.

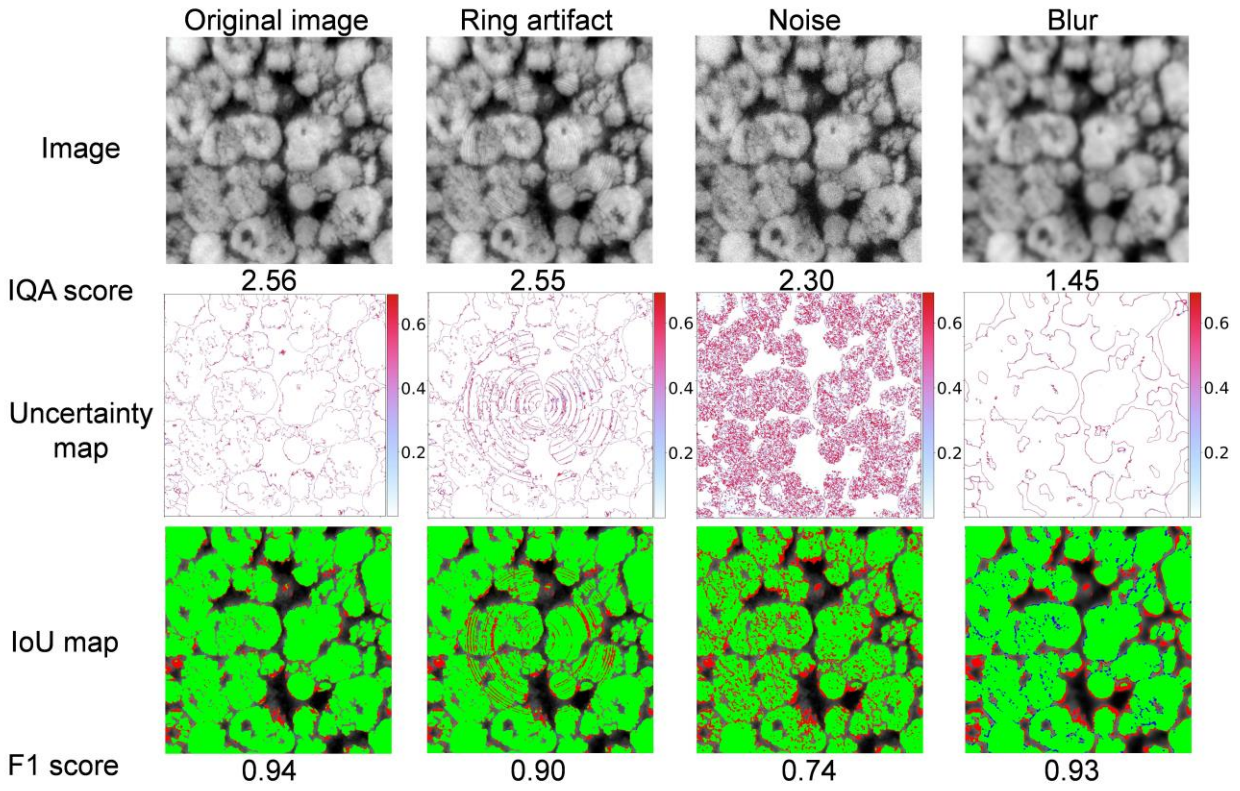


Figure 6 Results of different distorted images evaluated by TIQA and segmentation. For F1 score, it is in the range of 0 (the worst) and 1 (the best).

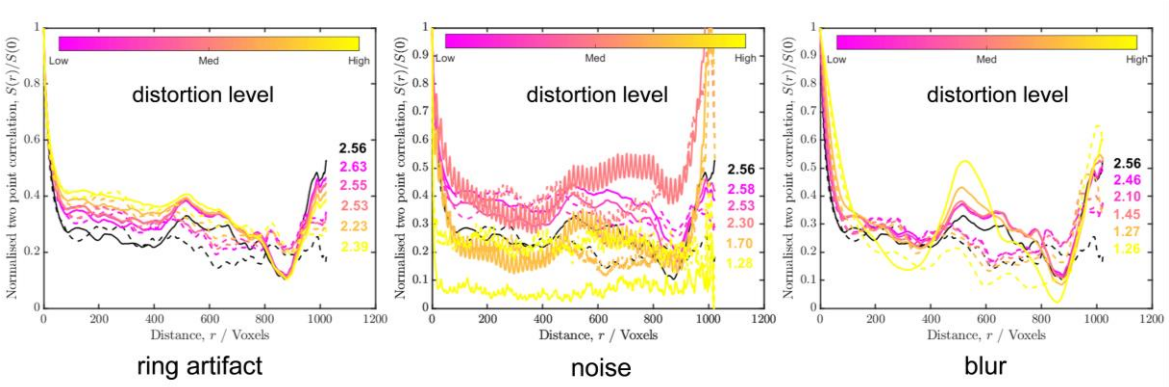


Figure 7 Point correlation between predicted segmentation mask and ground truth for black phase. The figures from left to right demonstrate the correlation results from images with different types of distortion. The color bar shows the distortion at different level, from little distortion to severe distortion. The solid line means the point correlation at X direction while the dash line indicates the relation at Y direction.

metric / distortion	SROCC	PLCC
Ring artifact	0.925	0.887
Noise	0.829	0.877
Blur	0.928	0.952

Table 3 Quantitative results of the correlation between predicted quality score and segmentation accuracy. These measurements are calculated between F1 score and IQA score.

Chapter 3 - Discussion

Tomography images are widely used for analyzing the material of the battery. However, the essential image pre-processing procedure is observer-dependence and misses a trust-worthy quantitative metric to select the image for reliable segmentation analysis. Although many DNN methods have been proposed for natural IQA, they need a huge amount of data for training, which is time-consuming and labor-intensive to create the dataset. Therefore, the IQA method required less or no training data is more suitable to apply on tomography images from scratch.

In this paper, we proposed a TIQA method including automatically generating the dataset and image quality score prediction. Besides, we explore the relation of image quality and segmentation performance. The qualitative and quantitative evaluation results prove that our method can well assess the image quality with different types of distortion, including blur, noise and ring artifact.

Compared with RankIQA, our method employed the more efficient feature extractor, EfficientNet, which achieves better performance for images with ring artifact and noise distortion, as shown in Table 2. Our network is also a fully convolutional neural network and could take the images regardless the size. Additionally, as a result of the data generation module, our method needs only dozens of training data, which is less than the RankIQA, about one-fifteenth of its total requirement. For the idea of using neural network to evaluate the results of IQA, we use the similar method as Samuel et al.[36], who investigated the influence of image quality on DNN results by applying different distorted images on the same network, but we conducted more types of distortion. Rather than use the classification result, the pixel-wise segmentation result is more concrete. Taking advantage of uncertainty map and IoU map, the influence of distortion can be visualized more comprehensively.

Despite the accurate results obtained by our network, some limitations also existed. Although we try to reduce the demand of training annotations, a small number of labels are still required and it cannot be regarded as an absolutely blind IQA method. Besides, the undistorted images are not well evaluated in our method, due to the lack of images with excellent quality.

Due to the demand of automatically analyzing the tomography images, our TIQA method can be extended to improve the image quality by using different image processing method. Our method can work as a teacher, which helps to train a DNN network to remove the distortion. It will greatly release the burden of human observers and reduce the impact of distortion on segmentation. In addition, the image quality assessment can be extended to object-oriented assessment. For example, through learning of object information, the network can judge whether the materials inside of the battery are destroyed or not.

In conclusion, we propose a TIQA method by integrating the FR-IQA method and NR-IQA method which achieves great performance with only a few

annotations for training. It greatly reduces the tedious work for selecting the good images and facilitates the automation of analyzing tomography images. In addition, it provides more reliable assessment on image pre-processing results, which avoids the conflicts of different human observers, and promises an outperformed segmentation analysis.

Chapter 4 – Methods

IV.1 Dataset creation

We collected 40 8-bit images from 11 different types of batteries with different resolutions. All the images were rescaled to the same resolution 224×224. To avoid changing the structure of some images, of which the width was not equal to the height, we resized the original image to the width or height equaling 255 and then randomly cropped the region with size of 224×224. We also maintained 6 original images for the analysis of the impact of down-sampling operation on image quality score. To expand the dataset, we have applied different algorithms with different parameters to generate images with different types of distortion. Totally, there are three common types of distortion: ring artifact, noise and blur. Similar to [37], we manually set the parameter values that control the distortion amount such that the visual quality of the distorted images varies, from an expected rating of 1(terrible) to 5 (excellent). The distortion parameter values were chosen based on a small set of images and applied the same for the remaining images in our database.

We performed two surveys for subjective image quality score and conveyed them to our colleagues for annotating a level from five levels, terrible, bad, average, good and excellent. Averagely, we collected 15 labels for each of the 111 images.

IV.2 Evaluation metrics

The PLCC is the linear correlation coefficient between the predicted score and human labeled score. It measures the prediction accuracy of an IQA metric, i.e., the capability of the metric to predict the subjective scores with low error. The PLCC is calculated as follows:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{M_d} \left(\hat{y}_i - \hat{y}_{avg}\right)\left(y_i - y_{avg}\right)}{\left(\sum_{i=1}^{M_d} \left(\hat{y}_i - \hat{y}_{avg}\right)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{M_d} \left(y_i - y_{avg}\right)^2\right)^{\frac{1}{2}}}$$

where $\hat{y}_i$ and y_i are the predicted score and the human labeled score of the *i-th* image in a dataset of size M_d respectively, $\hat{y}_{avg}$ and y_{avg} are the average of the predicted scores and human labeled scores respectively.

The SROCC is the rank correlation coefficient between predicted score and labeled score and it compares the monotonicity of the prediction performance, i.e., the limit to which the predicted scores agree with the relative magnitude of the labels. The SROCC can be calculated via following equation:

$$\mathrm{SROCC} = 1 - \frac{6\sum_{i=1}^{M_d}(d_i)^2}{M_d(M_d^2-1)}$$

where the d_i is the difference between the *i-th* image's rank in prediction results and labels.

The IoU and F1 score are utilized to measure the segmentation performance. IoU means the area of overlap between the predicted segmentation and the ground truth divided by the area of union between them. It ranges from 0 to 1 with 0 signifying no overlapping and 1 indicating perfect overlapping. Different from IoU, the F1 score can be calculated by:

$$\mathrm{F1} = \frac{2\times\ overlap}{total\ pixels}$$

where the total pixels mean the number of pixels in both segmentation results and ground truth.

IV.3 Data generation

As illustrated in Figure 2, the preprocessed images are regarded as reference image. Then several distortion filters, including noise, blur, and ring artifact, are applied on the reference image to generate the distorted images. The parameter values of the filters were set differently to create different distortion masks before adding them to the reference images so as to produce the images at different level of distortion. For label projection, we use five FR-IQA evaluators, mimicking the human observers, to calculate the difference between reference image and distorted image and donate a score for distorted image. Due to the range of the score from each evaluator varies, we normalized and rescaled them to the same range. Finally, we averaged the produced scores and set it as the generated score.

IV.4 Score prediction

As shown in Figure 1, we took the Efficientnet network as the feature extractor and change the last three layers to output a score for each input image. Among the dense layers, we added dropout[38] to avoid the overfitting. Rather than train the network from scratch, we transferred the weights from the pre-trained model in ImageNet[39] to reduce the time of convergence[40]. The input image size was fixed at 224×224×3 and the corresponding output was a score with shape of 1×1.

We built the image pair by picking an original image, generating the distorted images with distortions at different levels. The image with lower level of distortion is regarded as better image than the one with higher level of distortion. Taking advantage of the generated ranking information, the network can learn to order the images by quality. The corresponding rank loss[41] function is

$$\mathrm{L}(\hat{\mathrm{y}}_\mathrm{i},\hat{\mathrm{y}}_\mathrm{j}) = \max(0,\mathrm{m}+\hat{\mathrm{y}}_\mathrm{i}-\hat{\mathrm{y}}_\mathrm{j})$$

where $\hat{y}_i,\hat{y}_j$ are the prediction results of a pair of images, *m* is a margin to control the distance of the image pair.

After the image ordering process, the human annotations and the generated machine labels are inputted into the network to regress the output score to a fixed range by leveraging the Minor Square Error (MSE) loss function.

IV.5 Training and testing parameters

In score prediction module, we used 32 original images which were expanded to 512 images after data generation but without labels for training the rank. The initial learning rate was set at 3e-5 and decayed after several iterations. The network was trained for 30 epochs and on each epoch, it iterated on the whole dataset. The rate of the dropout was set at 0.5 to avoid overfitting. The Adam[42] optimizer was applied for optimizing the rank loss.

After training the rank, the model was finetuned in the score regression step. The training dataset contains 29 images with the size of 224*224*3 and their corresponding labels, which are in the range from 1 to 5. The data generation method was also implemented to expand the training dataset to 464 images with generated annotations. Then, they were inputted to the network for regression with the MSE loss. The network iterated 20 epochs with the initial learning rate at 5e-5 which decayed every 4 epochs. The dropout rate was 0.5 in training. For testing procedure, totally 56 images were tested and evaluated with corresponding human annotations.

All the experiments were conducted on python with TensorFlow [43] library. The computing hardware was Tesla K80.

IV.6 Segmentation-based evaluation method

To inspect the effect of distortion on segmentation accuracy, we applied D-LinkNet [44], which is a encoder-decoder network connected by dilated convolution [45], for tomography image segmentation (More details refer to Supplementary Figure S3). As presented in Figure 5, the network was trained on tomography images and annotated segmentation labels before making predictions. It segmented the image to 2 classes and produced the probability map which indicated the possibility of each pixel belonging to a class. The uncertainty map was generated by calculating the entropy of the possibility of each pixel belonging to the different classes. It represents the certitude of the network on the prediction results. The grey-scale maps show the “certainty” with which the generator assigns a phase to each voxel. A high certainty is represented as a white voxel, while a low certainty is represented as a red voxel. A higher uncertainty exists at the interphases, while low uncertainty exists at the bulk. The binary classification mask was produced by binarizing the possibility map, which is outputted from the network. The original image is finally classified into two classes, the phase 1 and phase 2 in our experiment, which can be extended to multiple phases. For the (Intersection-over-union) IoU map, it is created by comparing the binarized segmentation result and the ground truth map. The correctly predicted regions are marked green, the unsuccessfully predicted regions are marked blue and the wrongly predicted regions are colored

red.

The network ran for 200 epochs on 110 images with segmentation labels. The size of input image and label was 1024×1024 and they were normalized to range of 0 and 1 before inputted to the network. Instead of training from scratch, the encoder module was transferred from ImageNet pre-trained ResNet34 [46] model. The initial learning rate was 1e-4 and it decayed after fixed steps. The optimizer was Adam. Dice bice loss and binary cross-entropy loss were used to measure the difference between prediction and ground truth, and update the weights of the network.

In testing procedure, the output of the network was utilized to generate the uncertainty map. We use the entropy function [47] to calculate the uncertainty, which is described as follows,

$$H[y|x,X,Y] = -\sum_{c} p(y=c|x,X,Y)\log p(y=c|x,X,Y)$$

where x is the test image, y is the predicted class, X and Y are the images and labels in training process, *c* is the class index.

For the IoU map, we compared the binarized segmentation results with the labels and different areas were marked with different colors.

Chapter 5 – Conclusion

This paper proposed a method for TIQA, which requires a small number of human annotations but achieves excellent performance. This TIQA method is consisted of two modules: the data generation module expands the training dataset which addresses the problem of the lack of data; the score prediction module assesses the image quality and quantifies the quality score. With the estimated results, people can select the propriate image pre-processing method to generate a high-quality image.

Moreover, the relation between the predicted quality score and segmentation accuracy was investigated. The experiment results indicated that the segmentation was sensitive to some specific type of distortion, the noise distortion. It also revealed the relation between the quality score and segmentation performance, which provided a fast estimation of the performance of segmentation by inspecting the quality score.

For further usage of the data and algorithms, the readers can visit the project repository at https://github.com/SummerOf15/TomoIQA.

Reference

[1] P. Pietsch and V. Wood, “X-Ray Tomography for Lithium Ion Battery Research: A Practical Guide,” *Annu. Rev. Mater. Res.*, vol. 47, no. 1, pp. 451–479, Jul. 2017, doi: 10.1146/annurev-matsci-070616-123957.

[2] R. A. Ketcham and W. D. Carlson, “Acquisition, optimization and interpretation of x-ray computed tomographic imagery: Applications to the geosciences,” *Comput. Geosci.*, vol. 27, no. 4, pp. 381–400, May 2001, doi: 10.1016/S0098-3004(00)00116-3.

[3] S. Schlüter, A. Sheppard, K. Brown, and D. Wildenschild, “Image processing of multiphase images obtained via X-ray microtomography: A review,” *Water Resources Research*, vol. 50, no. 4. American Geophysical Union, pp. 3615–3639, Apr. 01, 2014, doi: 10.1002/2014WR015256.

[4] N. Otsu, “Threshold Selection Method From Gray-Level Histograms.,” *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, doi: 10.1109/tsmc.1979.4310076.

[5] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, Jan. 1984, doi: 10.1016/0098-3004(84)90020-7.

[6] M. J. Willemink and P. B. Noël, “The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence,” *Eur. Radiol.*, vol. 29, no. 5, pp. 2185–2195, May 2019, doi: 10.1007/s00330-018-5810-7.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[8] L. Zhang, Y. Shen, and H. Li, “VSI: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, 2014, doi: 10.1109/TIP.2014.2346028.

[9] W. Zhu, G. Zhai, Y. Liu, N. Lin, and X. Yang, “Reduced-reference image quality assessment based on free-energy principle with multi-channel decomposition,” Nov. 2018, doi: 10.1109/MMSP.2018.8547054.

[10] S. Golestaneh and L. J. Karam, “Reduced-Reference Quality Assessment Based on the Entropy of DWT Coefficients of Locally Weighted Gradient Magnitudes,” *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, Nov. 2016, doi: 10.1109/TIP.2016.2601821.

[11] W. S. Geisler, “Visual Perception and the Statistical Properties of Natural Scenes,” *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 167–192, Jan. 2008, doi: 10.1146/annurev.psych.58.110405.085632.

[12] A. Leclaire and L. Moisan, “No-Reference Image Quality Assessment and Blind Deblurring with Sharpness Metrics Exploiting Fourier Phase Information,” *J. Math. Imaging Vis.*, vol. 52, no. 1, pp. 145–172, May 2015, doi: 10.1007/s10851-015-0560-5.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, 2003, vol. 2, pp. 1398–1402, doi: 10.1109/acssc.2003.1292216.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, doi: 10.1109/TIP.2011.2109730.

[15] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 668–695, Feb. 2014, doi: 10.1109/TIP.2013.2293423.

[16] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, “Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator,” *IEEE Access*, vol. 4, pp. 5579–5590, 2016, doi: 10.1109/ACCESS.2016.2604042.

[17] M.-J. Chen and A. C. Bovik, “No-reference image blur assessment using multiscale gradient,” *EURASIP J. Image Video Process.*, vol. 2011, no. 1, pp. 1–11, Dec. 2011, doi: 10.1186/1687-5281-2011-3.

[18] A. Ciancio, A. L. N. T. Da Costa, E. A. B. Da Silva, A. Said, R. Samadani, and P. Obrador, “No-reference blur assessment of digital pictures based on multifeature classifiers,” *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011, doi: 10.1109/TIP.2010.2053549.

[19] N. D. Narvekar and L. J. Karam, “A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD),” *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Mar. 2011, doi: 10.1109/tip.2011.2131660.

[20] Q. Yan, Y. Xu, and X. Yang, “No-reference image blur assessment based on gradient profile sharpness,” 2013, doi: 10.1109/BMSB.2013.6621727.

[21] R. A. Manap and L. Shao, “Non-distortion-specific no-reference image quality assessment: A survey,” *Inf. Sci. (Ny).*, vol. 301, pp. 141–160, Apr. 2015, doi: 10.1016/j.ins.2014.12.055.

[22] W. S. Geisler, “Visual perception and the statistical properties of natural scenes,” *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 167–192, Jan. 2008, doi: 10.1146/annurev.psych.58.110405.085632.

[23] M. A. Saad, A. C. Bovik, and C. Charrier, “DCT statistics model-based blind image quality assessment,” in *Proceedings - International Conference on Image Processing, ICIP*, 2011, pp. 3093–3096, doi: 10.1109/ICIP.2011.6116319.

[24] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011, doi: 10.1109/TIP.2011.2147325.

[25] A. Mittal, A. K. Moorthy, and A. C. Bovik, “Blind/referenceless image spatial quality evaluator,” in *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 2011, pp. 723–727, doi: 10.1109/ACSSC.2011.6190099.

[26] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Sep. 2014, pp. 1733–1740, doi: 10.1109/CVPR.2014.224.

[27] K. Gu, G. Zhai, X. Yang, and W. Zhang, “Deep learning network for blind image quality assessment,” in *2014 IEEE International Conference on Image Processing, ICIP 2014*, Jan. 2014, pp. 511–515, doi: 10.1109/ICIP.2014.7025102.

[28] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, “Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.

[29] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, “On the use of deep learning for blind image quality assessment,” *Signal, Image Video Process.*, vol. 12, no. 2, pp. 355–362, Feb. 2018, doi: 10.1007/s11760-017-1166-8.

[30] X. Liu, J. Van De Weijer, and A. D. Bagdanov, “RankIQA: Learning from Rankings for No-Reference Image Quality Assessment,” in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2017, vol. 2017-Octob, pp. 1040–1049, doi: 10.1109/ICCV.2017.118.

[31] K. Y. Lin and G. Wang, “Hallucinated-IQA: No-Reference Image Quality

Assessment via Adversarial Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 732–741, doi: 10.1109/CVPR.2018.00083.

[32] M. Ebner, D.-W. Chung, R. E. García, and V. Wood, "Tortuosity Anisotropy in Lithium-Ion Battery Electrodes," *Adv. Energy Mater.*, vol. 4, no. 5, p. 1301278, Apr. 2014, doi: 10.1002/aenm.201301278.

[33] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Aug. 12, 2020. [Online]. Available: http://arxiv.org/abs/1905.11946.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, Accessed: Aug. 12, 2020. [Online]. Available: http://www.robots.ox.ac.uk/.

[35] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning and Data Mining*, Springer US, 2017, pp. 260–260.

[36] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016*, Jun. 2016, p. 7498955, doi: 10.1109/QoMEX.2016.7498955.

[37] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," 2019, doi: 10.1109/QoMEX.2019.8743252.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, 2014.

[39] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009, doi: 10.1109/cvprw.2009.5206848.

[40] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet Pre-training," no. i, pp. 1–10, 2018, [Online]. Available: http://arxiv.org/abs/1811.08883.

[41] W. Chen, T. Y. Liu, Y. Lan, Z. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," 2009.

[42] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, 2015.

[43] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," 2016.

[44] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 192–196, 2018, doi: 10.1109/CVPRW.2018.00034.

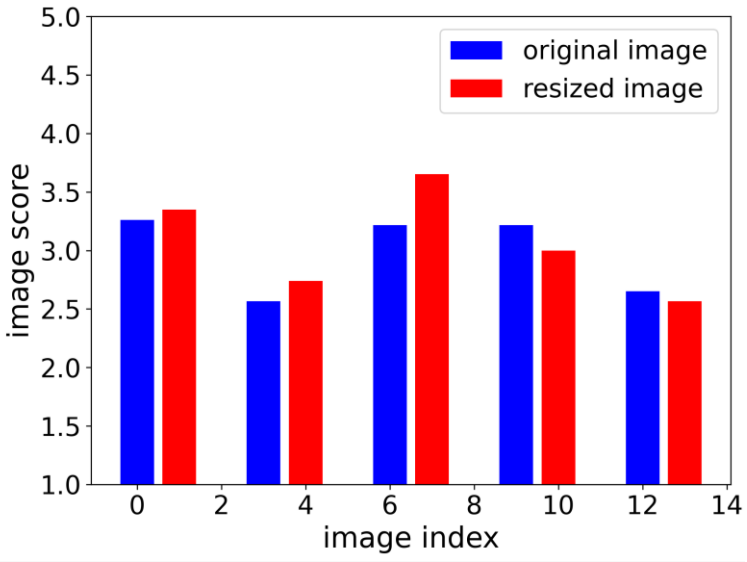
[45] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, [Online]. Available: http://arxiv.org/abs/1706.05587.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[47] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Annexes

A.1 Comparison IQA results of images at different scale



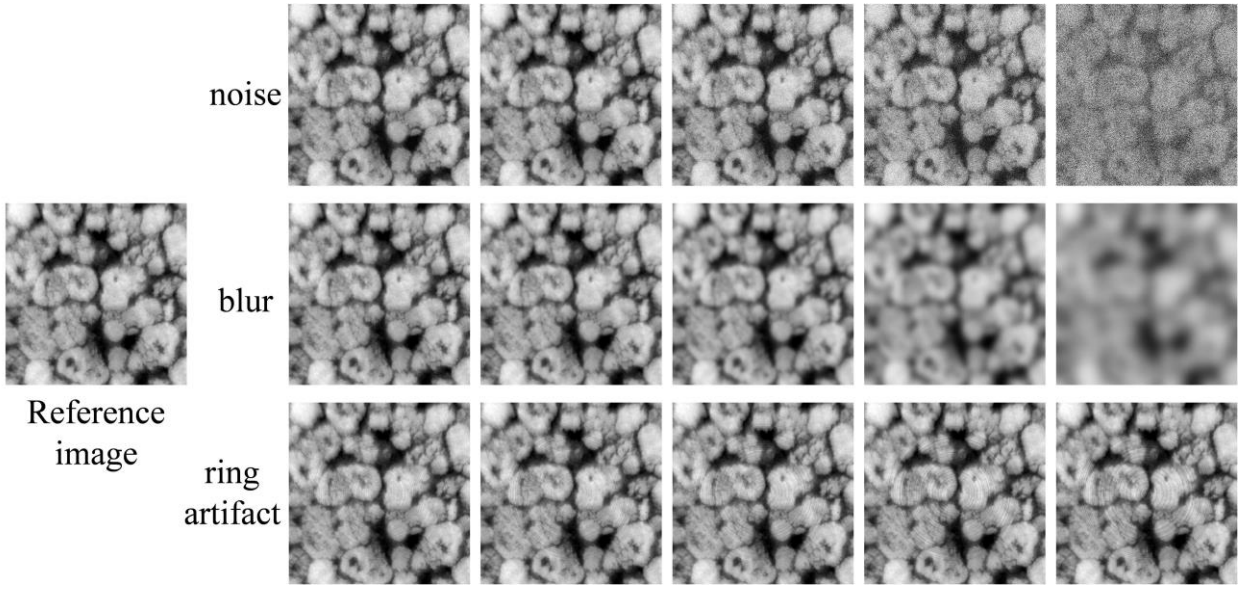
Supplementary Figure S1 Comparison of the labels for images before and after down-sampling.

A.2 Distorted image samples

As shown in Supplementary figure S2, the distorted images are generated based on reference image with different coefficients, which is listed in Supplementary Table 1. The blur and noise distortion are generated by adding the filters with different variance. For the ring artifact, the rings with different radius are added to original images. The rings have different pixel values at different level, which are proportional to the pixel values of original images.

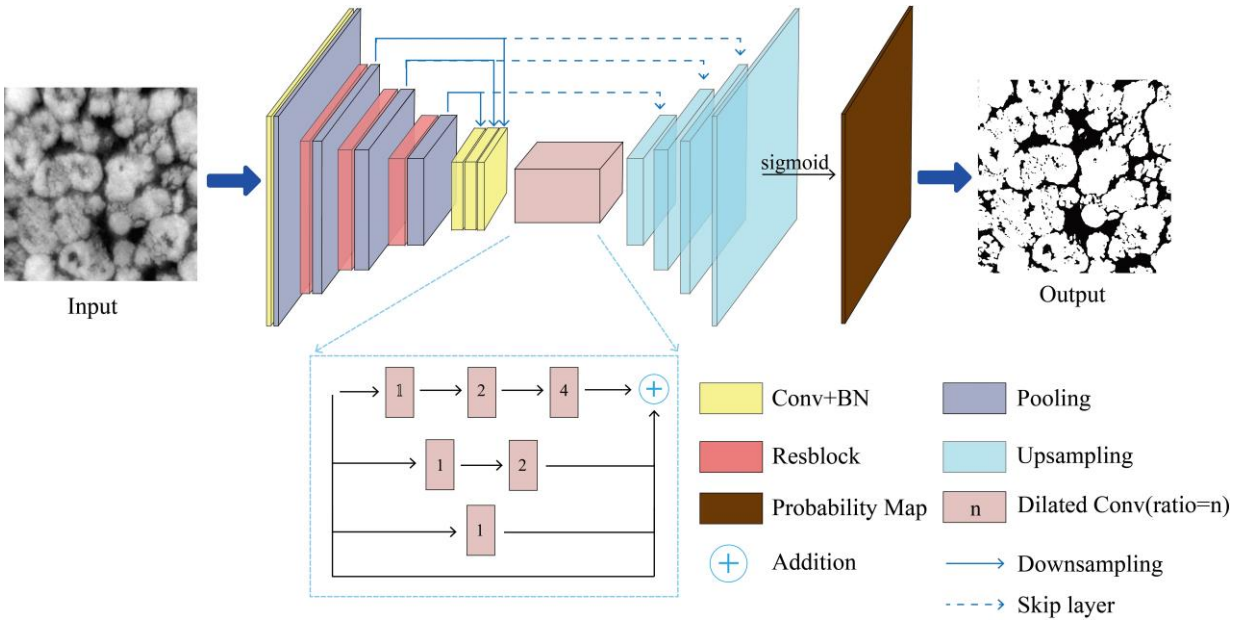
Distortion	**Variance (from level 1 to level 5)**
blur	1.2, 2.5, 6.5, 15.2, 33.2
noise	0.001, 0.006, 0.022, 0.088, 1.00
Ring artifact	0.1,0.13,0.15,0.18,0.2

Supplementary Table S1 Coefficients of generating distorted images.



Supplementary Figure S2 Results of distorted image generation. These three types of distorted images are produced from the reference image.

A.3 Segmentation network



Supplementary Figure S3 Architecture of segmentation network.

As shown in Figure S3, the segmentation network contains three modules: (a) the encoder module, which is based on the ResNet34; (b) the feature integration module, which concatenate these feature maps both in series and in parallel; (c) the decoder module, which is made up of deconvolutional layers to recover the feature map to the segmentation mask. In the training procedure, the encoder was initialized by the pre-trained ResNet34 weights. The output feature map of each encoding layer was down-sample to the same size as the last output of encoder. Then, these layers were integrated and inputted to the integration module. Dilated convolution was applied to increase the perspective filed of convolution operator without the lost of feature information. Through series and parallel connection, the semantic features were added

before inputted to the decoder module. The skip layer was utilized to transfer the low-level semantic feature (edges, corners) to the decoder module for enhancing the edge information of final segmentation mask. The sigmoid function is targeted to convert the multiple segmentation features to the possibility map. Finally, a threshold is selected to binarize the possibility map and produce the segmentation mask.

Glossary

CNNs Convolutional neural networks

DNNs Deep neural networks

FR-IQA Full-reference image quality assessment

HVS Human vision system

IoU Intersection to Union. IoU and F1 score are metrics for evaluating segmentation accuracy.

IQA Image quality assessment

LIBs Lithium ion batteries

MSE Mean squared error

NR-IQA No-reference image quality assessment

RR-IQA Reduced-reference image quality assessment

TIQA Tomography image quality assessment