LETTER

SEM Image Quality Assessment Based on Texture Inpainting

Zhaolin LU[†], Ziyang ZHANG^{†a)}, Yi WANG^{††}, Liang DONG^{†††}, Nonmembers, and Song LIANG[†], Member

SUMMARY This letter presents an image quality assessment (IQA) metric for scanning electron microscopy (SEM) images based on texture inpainting. Inspired by the observation that the texture information of SEM images is quite sensitive to distortions, a texture inpainting network is first trained to extract texture features. Then the weights of the trained texture inpainting network are transferred to the IQA network to help it learn an effective texture representation of the distorted image. Finally, supervised fine-tuning is conducted on the IQA network to predict the image quality score. Experimental results on the SEM image quality dataset demonstrate the advantages of the presented method.

key words: image quality assessment, no-reference, SEM image, texture-inpainting

1. Introduction

High-quality scanning electron microscopy (SEM) images are essential for studying the microscopic world [1]. However, the visual fatigue caused by long-term shooting can easily lead to the operator's imperfect adjustment of the imaging process, resulting in different types and degrees of distortions in the captured images. Therefore, effective SEM image quality assessment (IQA) metrics are needed to select high-quality images for subsequent analysis. When evaluating SEM image quality, the pristine reference image is usually not available, and the distortion type cannot be pre-determined, so the general purpose no-reference (NR) metrics are best suited for the SEM-IQA task.

General-purpose NR-IQA metrics can be divided into traditional metrics and deep learning-based metrics. Traditional metrics use hand-crafted features to measure image quality [2]–[6]. In contrast, deep learning-based metrics leverage the powerful expression capabilities of convolutional neural networks (CNN) to learn image quality features automatically and are generally more effective [7]. Kang et al. [8] and Bosse et al. [9] proposed end-to-end IQA models using CNN. Other improved methods, such as combining saliency detection with CNN [10] or predicting false

(a) Noise, 1.375 (b) Blur, 1.958 (c) Bright, 3.095 (d) Contrast, 4.167

Fig. 1 SEM images affected by four types of distortions. The subtitle indicates the distortion type and mean opinion score (MOS).

reference images [11], have also been proposed. However, None of those mentioned above aimed at SEM-IQA tasks. Recently, Li et al. [12] and Wang et al. [13] studied the quality evaluation of SEM images, but their works only covered the defocus blur distortion.

Due to the ultra-high magnification of electronic imaging and non-local self-similarity of the material microstructure, SEM images are mainly composed of repeated textures. Figure 1 shows SEM images degraded by four types of distortions. It can be seen that the fidelity and intelligibility of the hexagonal textures are quite sensitive to distortions. Therefore, we believe that the key to high SEM-IQA performance lies in the effective extraction of texture features.

To this end, we propose an SEM-IQA metric based on texture inpainting. The main idea is to utilize a texture inpainting pretext-task [14] to help the IQA target task learn better image texture representation. First, we train an encoder-decoder-shaped texture inpainting network, which extracts texture representation from the context to restore the texture information of the missing region. After training, the encoder has learned to extract texture features effectively, and due to the non-local self-similarity of SEM image texture, the extracted features can characterize the entire image. Then, we transfer the encoder's weight to the IQA network's down-sampling part to obtain an effective texture representation of the distorted input image. Finally, we perform supervised fine-tuning on the IQA network to predict the quality score.

2. Proposed Method

The overall framework of our SEM-IQA metric is shown in Fig. 2. It consists of a texture inpainting network and an IQA network. To train the texture inpainting network, we need to prepare specific edge-aware context and ground truth missing region. In the following sections, we first introduce the data generation for texture inpainting and then present the

Manuscript received September 18, 2020.

Manuscript revised October 18, 2020.

Manuscript publicized October 30, 2020.

[†]The authors are with School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China.

^{††}The author is with School of Engineering, Jiangsu Normal University Kewen College, Xuzhou 221116, China.

^{†††}The author is with School of Chemical Engineering and Technology, China University of Mining and Technology, Xuzhou 221116, China.

a) E-mail: zhangziyan@cumt.edu.cn

DOI: 10.1587/transinf.2020EDL8123

Fig. 2 Flowchart of the proposed method.

Fig. 3 Data generation for the texture inpainting network. The ratio of the missing region is changed for better visualization.

detailed implementation of the two networks.

2.1 Data Generation for Texture Inpainting

Unlike previous inpainting tasks that predict the missing region using context from the entire image [15], we add the missing region's edge map to the context, and generate the edge-aware context [16] as network input, as shown in Fig. 3. This modification makes the network focus on the inpainting of texture information, thus extracting texture features more effectively. We adopt the trained holistically-nested edge detection (HED) [17] model and non-maximum suppression (NMS) [18] to sequentially obtain the rough edge map and refined edge map of the distorted image. Then, the centered green box is utilized to extract the ground truth missing region, and the concentric smaller magenta box is employed to generate the rough and refined edge-aware context, forming two training sets. The narrow-band overlapping area of width l_{ol} between the two boxes appears in both the input and ground truth of the network and serves as a guide for the inpainting task. Besides, we leverage training set 1 and training set 2 to train the network successively because the rough edge-aware context contains more information and is used as a transition to make training more stable.

2.2 Texture Inpainting Network

2.2.1 Architecture

The texture inpainting network is an encoder-decoder

pipeline. The encoder contains seven down-sampling convolutional (denoted as conv) layers with 64, 64, 128, 256, 512, 512, and 1024 output channels. The decoder consists of five up-sampling convolutional (denoted as deconv) layers with 512, 256, 128, 64, and 1 output channel. All convolutional layers are with kernel size four. The second to the sixth conv layers use batch normalization, and a leakyrelu layer with a slope of 0.2 is adopted to activate all conv layers. The encoder is followed by a dropout layer with a retention rate of 0.5. The first four upconv layers are batch normalized and activated by relu layers. Except that the last conv layer and the first deconv layer are with stride one and padding zero, all other layers are with stride two and padding one.

2.2.2 Loss Function

Like context encoder [15], a reconstruction (L2) loss and an adversarial loss are adopted to handle the similarity in the overall structure and naturalness of the generated missing region. We modify the reconstruction loss by defining a weight matrix $\hat{W}$ to impose a greater penalty w_{ol} on the error in the overlapping region,

$$\hat{W}(i, j) = \begin{cases} 1 & i \in (l_{ol}, H - l_{ol}), j \in (l_{ol}, W - l_{ol}) \\ w_{ol} & \text{otherwise} \end{cases} \quad (1)$$

Where H, W are the height and width of the error map, and i, j denote the pixel indexes of the weight matrix. The modified reconstruction loss is the dot product of the L2 distance error map and the weight matrix,

$$\mathcal{L}_{rec} = \frac{1}{H \times W} (\hat{W} \odot \|I_{gt} - I_{pred}\|_2^2) \quad (2)$$

Where I_{gt} and I_{pred} represent the ground truth missing region and the predicted missing region, respectively.

The discriminator consists of six down-sampling convolutional layers and a sigmoid layer. All convolutional layers are with kernel size four and stride two, and the output channel numbers are 64, 128, 256, 512, 512, and 1. The first five layers are activated by a leakyrelu layer with a slope of 0.2, and the second to fifth layers are batch normalized. The discriminator takes in both I_{gt} and I_{pred} and tries to distinguish them, so the adversarial loss is:

$$\mathcal{L}_{adv} = \max_D \mathbb{E}[\log D(I_{gt}) + \log(1 - D(I_{pred}))] \quad (3)$$

We define the ratio of the reconstruction loss as λ_{rec} , then the final loss of the texture inpainting network is:

$$\mathcal{L}_{inpainting} = \lambda_{rec} \mathcal{L}_{rec} + (1 - \lambda_{rec}) \mathcal{L}_{adv} \quad (4)$$

2.3 Quality Prediction Network

The IQA network consists of a down-sampling part and a pooling part. For weight transfer, the down-sampling part is designed to be precisely the same as the first five layers of

the texture inpainting network. The pooling part includes a global average pooling (GAP) layer and two fully connected (FC) layers with output nodes of 16 and 1, respectively. The first FC layer is followed by a dropout layer with a retention rate of 0.5 to prevent overfitting.

Distorted images are cropped into patches to train the IQA network, and the quality label of the original image is adopted as the quality label of its patches. Given N training patches, we denote the prediction and ground truth quality score of the i -th patch as $\hat{y}_i$ and y_i . The squared Euclidean (L2) distance is employed as the loss function of the IQA network:

$$\mathcal{L}_{IQA} = \frac{1}{N} \sum_{i=1}^N \|\hat{y}_i - y_i\|_2^2 \quad (5)$$

When testing, the average prediction score of all image patches is adopted as the prediction score of the entire image.

3. Experimental Results

3.1 Dataset and Training Details

The SEM image quality dataset [12], [13] used in our experiments consists of 650 images with a resolution of 1024×884 and the corresponding mean opinion scores (MOS). It covers 50 samples in fields such as biology, chemistry, and mineralogy. Three noise-distorted images, three defocus blurred images, two brightness-distorted images, two contrast-distorted images, two astigmatism-distorted images, and one relatively high-quality image is taken for each sample. Images are classified into two categories according to whether it contains a specific background area, and the category sampling method is adopted to divide 70% of the data into the training set and 30% into the test set. The texture inpainting network and the IQA network share the same data division, while the former does not require MOS labels.

For both tasks, images are cropped into patches of size 256×256 , which is the best setting found experimentally. For the texture inpainting network, the HED algorithm and NMS post-processing are performed on the full-size image, while the subsequent stitching operation is conducted patchwisely. The size of the missing region is 64×64 , and $l_{ol} = 6$, $w_{ol} = 10$, $\lambda_{rec} = 0.998$. First, each SEM image and rough edge map are cropped into 12 non-overlapping patches to generate training set 1, then each SEM image and refined edge map are cut into 456 patches with a stride of 32 to form training set 2. Both the generator and discriminator are trained using Adam optimizer with a learning rate of $2e-4$ and a batch size of 220. We trained for 3000 epochs using training set 1, and then 30 epochs using training set 2. For the IQA network, we crop each SEM image into 16 patches and obtain a total of 7280 patches for training and 3120 patches for testing. After weight transfer, an Adam optimizer with an initial learning rate of $5e-4$ is used to fine-tune the IQA network with a batch size of 256. The learning

Table 1 Performance comparison on the SEM image quality dataset. Top performances are marked boldfaced.

Metrics	SRCC $\uparrow$	PLCC $\uparrow$	RMSE $\downarrow$
BIQI [2]	0.7521	0.7641	0.4932
BRISQUE [3]	0.7494	0.7299	0.5601
DESIQUE [4]	0.6938	0.7038	0.5431
NIQE [5]	0.6316	0.6554	0.5775
QAC [6]	0.4298	0.4485	0.6833
CNN [8]	0.8927	0.8770	0.3674
SDCNN [10]	0.8723	0.8625	0.3870
DNN [9]	0.8733	0.8490	0.4040
RAN4IQA [11]	0.8357	0.5494	0.6389
Proposed	0.9148	0.8922	0.3453

rate is reduced ten-fold when training loss does not provide a new minimum for 15 consecutive epochs, and training stops when the learning rate is decreased the third time.

3.2 Performance Evaluation

Spearman's rank-order correlation coefficients (SRCC), Pearson linear correlation coefficients (PLCC), and root mean square error (RMSE) are adopted to evaluate the performance of the proposed metric [12]. The predicted quality scores are passed through a nonlinear logistic function before computing PLCC and RMSE [19]:

$$f(x) = \ell_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\ell_2(x - \ell_3)}} \right) + \ell_4 x + \ell_5 \quad (6)$$

where x denotes the score predicted by the objective assessment metric, $f(x)$ represents the mapped objective score, and $\ell_1, \ell_2, \ell_3, \ell_4, \ell_5$ are the fitting parameters.

3.2.1 Comparisons with the State-of-the-Arts

Our method is compared with four state-of-the-art deep learning-based metrics and five traditional metrics, namely CNN [8], SDCNN [10], DNN [9], RAN4IQA [11], BIQI [2], BRISQUE [3], DESIQUE [4], NIQE [5], and QAC [6]. We reduced the blocks of the evaluator module in RAN4IQA [11] from 5 to 3, and set its FC layer nodes to 256, 512, 16, and 1, to prevent overfitting for fair comparison, and the simplified module has a similar number of parameters to our IQA network. Table 1 shows the comparison results. As can be seen, the performance of the proposed metric outperforms other state-of-the-art metrics in terms of both prediction monotonicity (SRCC) and prediction accuracy (PLCC, RMSE). This result is consistent with the conjecture we have drawn from the characteristics of SEM images, i.e., the key to obtaining good SEM-IQA performance lies in the effective extraction of its texture features.

3.2.2 Performance Comparison on Blur Distortion

Defocus blur is the most common type of SEM image distortion, so our metric is further compared with two latest SEM-blur metrics [12], [13] and the above general-purpose metrics on the blur subset. As can be seen in Fig. 4, our

Fig. 4 Performance comparison on the defocus blur distortion. For better visualization, the worst-performing NIQE [5] and QAC [6] are omitted.

Table 2 Performance comparison of ablation study on the overall data set and various distortion types. The first two rows indicate SRCC performance, while the last two rows indicate PLCC performance.

Perf.	All	Noise	Cont.	Blur	Brig.	Asti.
Baseline	0.8931	0.8566	0.8676	0.8458	0.8293	0.7226
Proposed	0.9148	0.8896	0.8854	0.8764	0.8070	0.7131
Baseline	0.8701	0.8064	0.8308	0.8500	0.7899	0.7727
Proposed	0.8922	0.8778	0.8560	0.8603	0.8065	0.8049

metric still achieves the best results.

3.3 Ablation Study

3.3.1 Performance Gain

To further investigate whether the effectiveness of our approach is derived from weight transfer from the texture inpainting network, we conduct ablation studies. In the baseline method, the training data, training settings, and IQA network architecture are the same as the proposed method. However, the difference is that the weights of the IQA network are initialized using He initialization [20] instead of transferring from the texture inpainting network. He initialization is the most widely used weight initialization method and is the default initialization setting in PyTorch. We have also tried several other initialization methods, and the results are slightly worse than He initialization (e.g., the SRCC and PLCC performances obtained by training the network initialized by Xavier initialization [21] are 0.8897 and 0.8584, respectively). Table 2 shows the performance comparison of the ablation study. In terms of overall performance, our metric surpasses the baseline on both SRCC and PLCC. Besides, the SRCC performance for evaluating noise, contrast and blur distortion and the PLCC performance for evaluating all types of distortions are improved.

3.3.2 Impact of Training Samples

IQA is a typical small sample problem [7], but our self-supervised texture inpainting network can leverage extensive unlabeled samples to learn image features. In this section, we further illustrate our metric's low dependence on labeled data. The entire training set is still used to train the texture inpainting network, while we fix the test set but sample a decreasing training set for the IQA network. We randomly sample 11 times for each ratio and report the

Fig. 5 SRCC and PLCC performances of the proposed metric and baseline with a decreasing number of training samples. The horizontal axis represents the sampling ratio of the training set.

mean and standard deviation of the performance in Fig. 5. It can be seen that with weight transfer, our IQA network can achieve comparative performance using only 40% of the labeled data.

4. Conclusion

In this paper, an SEM-IQA metric based on texture inpainting is proposed. Based on the observation that the texture information of SEM images is quite sensitive to distortions, a texture inpainting network is adopted to help the IQA network learn more effective texture feature representation through weight transfer. Then supervised fine-tuning is conducted on the IQA network to obtain the quality score prediction. Experiments have proved that our metric outperforms state-of-the-arts and is less dependent on training samples.

Acknowledgments

This work was supported by the Major Scientific and Technological Innovation Projects in Shandong Province (2019JZZY020505), Special Emergency Project for Scientific Research on Novel Coronavirus (2019-nCoV) Prevention and Control (2020XGYJ10).

References

- [1] D. Chen, A. Miyamoto, and S. Kaneko, "Robust surface reconstruction in sem using two bse detectors," *IEICE Trans. Inf. & Syst.*, vol.E96-D, no.10, pp.2224–2234, 2013.
- [2] A.K. Moorthy and A.C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol.17, no.5, pp.513–516, 2010.
- [3] A. Mittal, A.K. Moorthy, and A.C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol.21, no.12, pp.4695–4708, 2012.
- [4] Y. Zhang and D.M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron. Imaging*, vol.22, no.4, p.043025, 2013.
- [5] A. Mittal, R. Soundararajan, and A.C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol.20, no.3, pp.209–212, 2012.
- [6] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.995–1002, 2013.
- [7] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A.C. Bovik,

- “Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment,” *IEEE Signal Process. Mag.*, vol.34, no.6, pp.130–141, 2017.
- [8] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1733–1740, 2014.
- [9] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, “A deep neural network for image quality assessment,” 2016 *IEEE Int. Conf. Image Process. (ICIP)*, pp.3773–3777, IEEE, 2016.
- [10] S. Jia and Y. Zhang, “Saliency-based deep convolutional neural network for no-reference image quality assessment,” *Multimed. Tools Appl.*, vol.77, no.12, pp.14859–14872, 2018.
- [11] H. Ren, D. Chen, and Y. Wang, “Ran4iqa: Restorative adversarial nets for no-reference image quality assessment,” *AAAI Conf. Artificial Intelligence*, pp.7308–7314, 2018.
- [12] Q. Li, L. Li, Z. Lu, Y. Zhou, and H. Zhu, “No-reference sharpness index for scanning electron microscopy images based on dark channel prior,” *KSII Trans. Internet & Inf. Syst.*, vol.13, no.5, 2019.
- [13] H. Wang, X. Hu, H. Xu, S. Li, and Z. Lu, “No-reference quality assessment method for blurriness of sem micrographs with multiple texture,” *Scanning*, vol.2019, 2019.
- [14] I. Misra and L.v.d. Maaten, “Self-supervised learning of pretext-invariant representations,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp.6707–6717, 2020.
- [15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A.A. Efros, “Context encoders: Feature learning by inpainting,” *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.2536–2544, 2016.
- [16] L. Liao, R. Hu, J. Xiao, and Z. Wang, “Edge-aware context encoder for image inpainting,” 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3156–3160, IEEE, 2018.
- [17] S. Xie and Z. Tu, “Holistically-nested edge detection,” *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp.1395–1403, 2015.
- [18] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.PAMI-8, no.6, pp.679–698, 1986.
- [19] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, and X. Gao, “No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness,” *IEEE Trans. Image Process.*, vol.28, no.9, pp.4566–4579, 2019.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp.1026–1034, 2015.
- [21] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, pp.249–256, 2010.